Psychometrika

CONTENTS

VOLUME EIGHT DECEMBER 1943 NUMBER FOUR

PSYCHOMETRIKA—VOL. 8, NO. 4
DECEMBER, 1943

A STATISTICAL TEST FOR MEANS OF SAMPLES FROM SKEW POPULATIONS

LEON FESTINGER
IOWA CHILD WELFARE RESEARCH STATION
STATE UNIVERSITY OF IOWA

This paper presents a test for determining significance of differences between means of samples which are drawn from positively skewed populations, more specifically, those having a Pearson Type III distribution function. The quantity $2npx_s/x_p$ (where p equals the mean squared divided by the variance and n is the number of cases in the sample), which distributes itself as Chi Square for $2np$ degrees of freedom, may be referred to the tables of Chi Square for testing hypotheses about the value of the true mean. For two independent samples, the larger mean divided by the smaller mean, which distributes itself as F for $2n_1p_1$ and $2n_2p_2$ degrees of freedom, may be referred to the F distribution tables for testing significance of difference between means. The test assumes that the range of possible scores is from zero to infinity. When a lower theoretical score limit (c) exists which is not zero, the quantity (Mean $-$ c) should be used instead of the mean in all calculations.

Exact tests of significance for means are known when the samples are drawn from normally distributed populations. In a previous article (2) a method has been described for testing significance of means when the samples come from exponentially distributed populations (J curves.) Between these two lies a range of skew frequency distributions for which we as yet have no statistical tests of significance. It is the purpose of this article to generalize the test for exponential distributions to all skew distributions of Pearson's Type III, of which the exponential distribution is a special case. The types of distributions to which this test may be applied can be seen in Fig. 1, where are plotted four Pearson Type III curves ranging from marked skewness to moderate skewness.

Derivations

Let us write the distribution function with which we shall deal in a general form

$$f(x) = cx^{p-1}\,e^{-bx}; \quad 0 \le x < \infty. \tag{1}$$

If we now set

$$c \int_0^\infty x^{p-1}\; e^{-bx}\, dx = k, \tag{2}$$

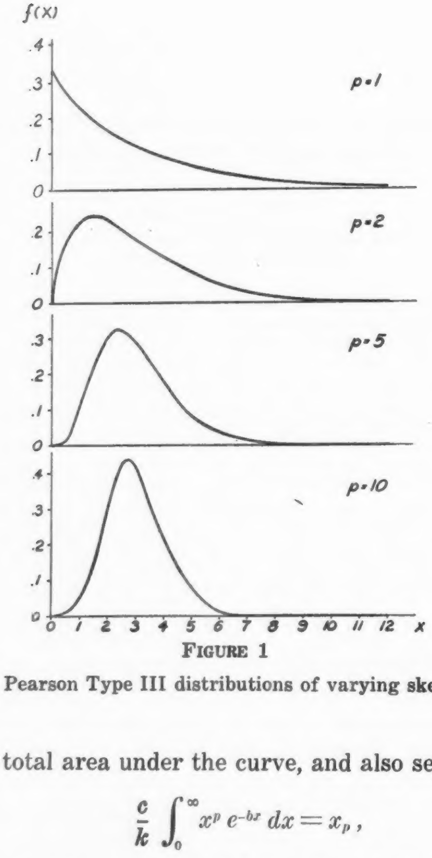

FIGURE 1

Examples of Pearson Type III distributions of varying skewness with $x = 3$.

where k is the total area under the curve, and also set

$$\frac{c}{k}\int_0^\infty x^p\, e^{-bx}\, dx = x_p\,, \tag{3}$$

where x_p is the arithmetic mean of the distribution, we find after evaluating the integral that

$$\frac{c\,\Gamma(p)}{b^p} = k \quad \text{and} \quad b = \frac{p}{x_p}.$$

We may therefore rewrite equation (1) so that the constants are expressed in terms of the moments of the distribution:

$$f(x) = \frac{k(p)^p}{x_p^p\,\Gamma(p)}\, x^{p-1}\, e^{-(px/x_p)}; \quad 0 \leq x < \infty: \tag{4}$$

Let us now set the area (k) equal to unity.

Craig (1) and Irwin (3) have both, by different methods, arrived at the distribution function of the means (x_s) of samples of size n drawn from the population of equation (4):

$$\phi(x_s) = \frac{(np)^{np}}{x_s{}^{np}\,\Gamma(np)}\, x_s{}^{np-1}\, e^{-(npx_s/x_p)}. \tag{5}$$

By rearrangement of terms we may write this in a more convenient form:

$$\phi(x_s) = \frac{\left(\frac{np}{x_p}\right)\left(\frac{npx_s}{x_p}\right)^{np-1} e^{-(npx_s/x_p)}}{\Gamma(np)}: \tag{5a}$$

Similarly to the procedure described in connection with the exponential populations (2), we may now reduce this to a Chi Square distribution by making the substitution

$$z = \frac{2\,n\,p\,x_s}{x_p}, \tag{6}$$

and of course

$$dx_s = \frac{x_p}{2\,n\,p}\, dz. \tag{7}$$

Then, setting

$$\int_{x_{s1}}^{x_{s2}} f(x_s)\,dx_s = \int_{(x_p/2np)z_1}^{(x_p/2np)z_2} f(z)\,dz\,, \tag{8}$$

we obtain

$$F(z) = \frac{\left(\frac{z}{2}\right)^{np-1} e^{-(z/2)}}{2\,\Gamma(np)} = F\left(\frac{2\,n\,p\,x_s}{x_p}\right) \tag{9}$$

which is the distribution of Chi Square for $2np$ degrees of freedom.

Thus the quantity $2npx_s/x_p$ may be referred to the table of Chi Square for the appropriate number of degrees of freedom for the purpose of determining the fiducial limits of the true mean for any obtained sample mean. This quantity, it may be observed, is identical with the quantity used for the exponential populations when p equals 1, under which circumstances equation (4) reduces to an exponential distribution. It is obvious that p is a measure of the skewness of the distribution; the larger p is, the less skew is the distribution.

It is interesting to note the different results obtained between this test and the application of a 't' test to distributions of this type.

This is shown in Fig. 2, where the 2 per cent fiducial limits have been calculated for the case of the obtained sample mean equal to 1, and the number of cases in the sample equal to 9 for p varying from 1 to 30.

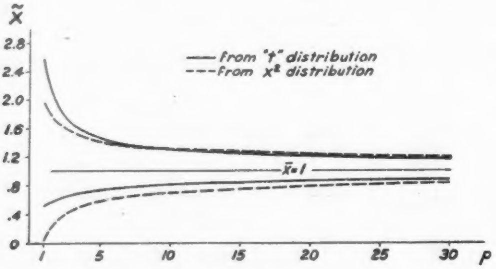

FIGURE 2

Comparison of 2 per cent fiducial limits determined by 't' test and Chi Square test for samples of $x = 1$, $n = 9$.

In order to apply the 't' test it is necessary to know the variance of the sample. If we set

$$\frac{c}{k}\int_0^\infty x^{p-1}\, e^{-bx}(x - x_s)^2\, dx = \sigma^2, \tag{10}$$

we find

$$\sigma^2 = \frac{1}{p}\, x_s^{\,2}. \tag{11}$$

Using the right-hand quantity of equation (11) we obtained the fiducial limits consequent to applying a 't' test as shown by the broken lines in Fig. 2. The solid lines show the fiducial limits obtained from referring $2npx_s/x_p$ to the Chi Square table. As is to be expected, the more skew the population, the greater the error introduced by the use of the 't' test. For p greater than 15, for this number of cases in the sample, there is little difference between the results obtained from the two tests. Thus, if the population is only slightly skew the 't' test can be used without appreciable error. For markedly skew populations, however, the 't' test is not permissible and the correct test described in this paper should be used.

For the test of significance between means we may apply the same reasoning which we applied previously (2). Since $2npx_s/x_p$ is distributed as Chi Square for $2np$ degrees of freedom, and since x_s

is an unbiased estimate of x_p, then, assuming independence of x_{s1} and x_{s2},

$$\frac{x_{s1}}{x_{s2}} = F \quad \text{(for } 2n_1p_1 \text{ and } 2n_2p_2 \text{ degrees of freedom)}. \qquad (12)$$

Thus the ratio of the larger to the smaller mean may be referred to the table of F for the appropriate number of degrees of freedom.

It is, of course, clear that before we can refer our statistics to either the table of Chi Square or the table of F with the appropriate number of degrees of freedom, we must determine the value of p. Referring back to equation (11), we see that p may be evaluated simply by dividing the mean squared by the variance. To be absolutely correct, the sampling distribution of p should be taken into account in our formulas. This is not, however, a serious objection since the sampling distributions of Mean and Sigma are positively correlated and the variance of the sampling distribution of p will, therefore, in any case, be quite small. Very little error will therefore be introduced by ignoring the sampling distribution of p.

We shall not present any concrete applications of the test at present, since several illustrations are given in the previous paper (2) and the method of analysis is identical. We may, however, summarize by reviewing the exact method of applying the present test.

Summary

The test applies to samples drawn from populations with a Pearson Type III frequency distribution. Most skew distributions obtained in practice will probably be adequately represented by a Pearson Type III curve. In cases of doubt a Chi Square distribution of $2p$ degrees of freedom may be fitted to the distribution of $2px/x_s$ and tested for goodness of fit.

When the test is applied to the difference between two means, the two samples are assumed to be independent.

If the lowest possible score is some constant 'c' greater than zero, the quantity used in application of the test should be $(x_s - c)$.

To find the fiducial limits of the true mean, the two values of Chi Square at the desired level of significance times the true mean are each in turn set equal to $2npx_s$.

To test the significance of difference between means, equation (12) is used. In calculating this ratio the larger mean should always be used as the numerator.

The value of p is determined by equation (11).

In cases where p is large, say over 10, and the number of cases in the samples is appreciable, say greater than 15, the 't' test may be used without much error.

REFERENCES

1. Craig, C. C. Sampling when the parent population is of Pearson's Type III. *Biometrika*, 1929, 21, 287-293.
2. Festinger, L. An exact test of significance for means of samples from populations with an exponential frequency distribution. *Psychometrika*, 1943, 8, 153-160.
3. Irwin, J. O. On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson's Type II. *Biometrika*, 1927, 19, 225-239.

PSYCHOMETRIKA—VOL. 8, NO. 4
DECEMBER, 1943

THE SIGNIFICANCE OF RANK DIFFERENCE COEFFICIENTS OF CORRELATION

G. R. THORNTON
PURDUE UNIVERSITY

The coefficients of rank difference correlation that are barely significant at six different levels of significance are given for *N*'s of 2 to 30. Most of the values were obtained by translation of Olds' tables of probabilities for various values of Σd^2. Comparison of these data with those obtained by four other methods indicates that one method yields values more appropriate than those obtained from Olds' data for coefficients significant at the .01 level for *N*'s from 11 to 25. This method also provides a convenient means of obtaining approximate values of coefficients significant at the .01 level for *N*'s above 30. Need for caution in evaluating the significance of coefficients obtained from data involving tie rankings is indicated. The article concludes with recommendations as to choice of methods of determining the significance of rank difference coefficients.

Until recently there has been no satisfactory method for testing the significance of a coefficient of correlation (r') obtained by Spearman's rank difference method by the formula,

$$r' = 1 - \frac{6\sum d^2}{N^3 - N}. \tag{1}$$

In a recent paper E. G. Olds (5), extending the work of Hotelling and Pabst (3), presents tables giving the probabilities for various values of $\sum d^2$ for *N*'s of 2 to 30. These tables should be very useful to persons who wish to test the significance of coefficients while in the process of calculating them by the formula given above.

A translation of Olds' tables into a table giving probabilities for the coefficients themselves seems desirable for two reasons: (1) Methods of calculating a rank difference coefficient of correlation are now in use which do not require the calculation of $\sum d^2$ (1). (2) Olds' tables are not readily applicable to the interpretation of rank difference coefficients reported in the past or present psychological literature.

Table 1 presents a translation of Olds' probabilities into rank difference coefficients of correlation which are barely significant at each of several levels of significance for *N*'s of 2 to 30. The probabilities listed at the head of the columns apply to both positive and

TABLE 1

Coefficients of Rank Difference Correlation that are Barely Significant at Given Levels of Significance

N	.01		.02	.04	.05†	.10	.20
2	none		none	none	none	none	none
3	none		none	none	none	none	none
4	none		none	none	none	.977	.883
5	none		.989	.948	.933	.861	.729
6	.960		.928	.868	.843	.774	.636
7	.906		.867	.806	.781	.695	.571
8	.869		.816	.749	.725	.634	.514
9	.825		.771	.705	.681	.592	.477
10	.788		.733	.668	.644	.555	.445
11	.816	(.764)*	.736	.651	.626	.521	.406
12	.777	(.736)	.702	.620	.599	.496	.387
13	.745	(.711)	.672	.594	.575	.476	.371
14	.720	(.687)	.646	.571	.553	.457	.356
15	.689	(.667)	.623	.550	.535	.441	.343
16	.666	(.648)	.602	.531	.517	.426	.332
17	.645	(.630)	.583	.514	.501	.412	.321
18	.626	(.614)	.565	.499	.487	.400	.312
19	.608	(.598)	.549	.485	.474	.389	.303
20	.592	(.583)	.535	.472	.462	.378	.295
21	.577	(.571)	.521	.460	.450	.369	.288
22	.563	(.558)	.509	.449	.440	.360	.281
23	.550	(.547)	.497	.439	.430	.352	.274
24	.538	(.536)	.486	.429	.420	.344	.268
25	.527	(.525)	.476	.420	.412	.337	.263
26	.516		.466	.412	.404	.330	.257
27	.506		.457	.404	.396	.324	.252
28	.497		.449	.396	.389	.318	.248
29	.488		.441	.389	.382	.312	.243
30	.479		.433	.382	.375	.306	.239

* The values given in parentheses are preferable. (See text below.)

† The values given for *N*'s of 11 to 30 in the .05 column are not so accurate as other values in the table. Their derivation is explained below.

negative coefficients. The coefficient .960 appearing opposite an N of 6 in the .01 column, for example, means that by chance a coefficient as large as .960 or larger may be expected to occur once in 200 times in the positive direction and once in 200 times in the negative direction. Where "none" appears in the table, it is to be interpreted to mean that no coefficient is significant at that level for the given N.

To obtain the coefficients in Table 1 from Olds' data several adjustments were necessary. For N's from 11 to 30 Olds presents limits of $\sum d^2$ within which 99 per cent, 98 per cent, etc., of the cases will occur by chance. To obtain the figures in Table 1 for N's from 11 to 30, inclusive, I have merely calculated the coefficients of correlation for the limits given by Olds and then added to each .001, with a few exceptions. The exceptions include the following: all coefficients for N of 12; the coefficient at the .01 level for N of 14; the coefficient at the .02 level for N of 29. In these instances there are minor errors in the data presented by Olds, for which I have attempted corrections before calculating the coefficients. Olds does not give the data from which to calculate the coefficients that are significant at the .05 level. The method of calculating these coefficients as given in Table 1 is explained later.

For N's of 2 to 10, inclusive, Olds presents probabilities for each of the possible values of $\sum d^2$, excluding data in which tie rankings occur. To determine the exact point at which the coefficient is barely significant at each of the levels of significance, it was necessary to interpolate. Interpolation was done by means of the normal curve. The resulting coefficients, given in Table 1, separate clearly the discrete significant and non-significant coefficients which are possible when data involving tie rankings are excluded. Any type of interpolation would have accomplished this. The value of interpolation lies in providing a dividing line between significant and non-significant coefficients which because of tie rankings fall between the highest non-significant and lowest significant coefficients possible without tie rankings. In the absence of evidence concerning the distribution of coefficients possible in data involving tie rankings, a normal distribution of them was assumed to be the best guess. Further consideration of this question is presented later.

Adequacy of Olds' Data for Testing Significance of Rank Difference Coefficients of Correlation

All statements in this section apply only to coefficients obtained from data in which tie rankings do not occur. With this limitation it is fair to say that Olds' (5) data provide the best basis yet available for testing the significance of rank difference coefficients of correlation.

The probabilities of various values of $\sum d^2$ which Olds presents for N's from two to seven, inclusive, are based upon tabulations of all possible combinations of ranks for each N. Kendall and others (4) have presented the results of similar tabulations which agree exactly with those of Olds. I have tabulated the possible combinations for N's from two to six, inclusive, and these results also agree perfectly with Olds' figures. Complete accuracy of Olds' probabilities for N's up to and including seven, therefore, may be assumed.

For N's of eight to ten, inclusive, Olds obtained approximate frequencies for various values of $\sum d^2$ by computations based upon type II curves. Evidence is presented to show that for purposes of determing levels of significance the approximations are very close (5). Kendall and others (4) present exact frequencies for N equals eight. I have compared the probabilities obtained from their frequencies with the probabilities calculated by Olds. There are slight differences, but the agreement is so close that in terms of dividing significant from non-significant coefficients for all of the levels listed in Table 1 of this report (still excluding from consideration all sets of data in which tie rankings occur) there are no disagreements. Olds' data, therefore, appear to be adequate for determining levels of significance for N's of eight, nine, and ten.

For N's from 11 to 30, inclusive, Olds has calculated probabilities on the basis of assumed normal curves. Both Olds (5) and Kendall and others (4) have presented evidence showing that the larger the N the more reasonable is the assumption of normal distribution. Olds' data indicate that while there are probably discrepancies between the exact distributions and his approximations for N's above ten, these differences are in all probability small. While admittedly not completely accurate, then, Olds' probabilities for N's above ten are the best available at present and for practical purposes are probably adequate in most instances.

Inspection of Table 1 of this report reveals some discrepancies inherent in Olds' data which are not apparent in Olds' tables in which tabulations are in terms of $\sum d^2$ only. The reader may spot these discrepancies by comparing the coefficients that are significant at the .01, .02 and .04 levels for N's of 10 and 11. It is surely unreasonable to assume that in order to be significant at the .01 or .02 level a coefficient found for an N of 11 should be higher than one found for an N of 10. Apparently the assumption of normal distribution has led to over-estimation of frequencies near the extremes of the distribution for an N of 11. It is to be noted that the distortion decreases as one moves in from the ends of the distribution and apparently becomes a distortion in the opposite direction near the point outside of which five per cent of the cases fall at one end of the distribution (the point represented by .10 level of significance in Table 1). It seems fair to assume that both of these distortions decrease as N increases.*

Comparison of Levels of Significance Obtained from Olds' Data with Those Obtained by Other Methods

The formula

$$\sigma_{r'} = \frac{1.05(1 - r'^2)}{\sqrt{N - 1}} \qquad (2)$$

has been suggested as a rough means of determining the reliability of a rank difference coefficient (2, p. 230). Now that we have empirical data or approximations to empirical data on the probabilities of various values of rho for N's of 2 to 30, it is possible to test the adequacy of this formula as a basis for calculating significance. Two methods of testing the usefulness of the formula have been employed: (1) calculation by means óf the formula of the t ratios for coefficients in Table 1 that are barely significant at the .01 and .05 levels; (2) calculation by means of the formula of the coefficients for selected N's that would be barely significant at the .01 and .05 levels according to the t ratio for the appropriate degrees of freedom ($N-2$). The resulting data are given in Tables 2 and 3. Comparison of these data with the expected t ratios and with the coefficients that are found to be significant by the empirical evidence from Olds indicates that the formula for the standard error of a rho is of little value in testing the significance of rho for N's of 30 and below. The inappropriateness of the formula is greatest for small N's and decreases as N increases. The formula results in an over-estimation of the significance of rhos for all N's up to 30. (Although the data for .05 level are given only for N's up to 10, these statements apply to both the .01 and .05 levels for N's up to 30. At the .05 level the overestimation is slight for N of 30.)

Since the above formula results in a too lenient test of significance, it was thought that perhaps the formula

* Comparison of the coefficient (.816) for N equals 11 in the column for .01 level of significance in Table 1 with the coefficient that one would obtain by extrapolation from the previous coefficients in the same column, indicates that the coefficient .816 is probably not in error by more than .06. This error appears to be the largest that occurs anywhere in Table 1.

$$\sigma_{r'} = \frac{1.05}{\sqrt{N-1}} \tag{3}$$

(assuming rho to be zero) would give better results. As the data in Tables 2 and 3 show, however, this formula results in a too conservative estimate of the significance of rhos for *N*'s up to 30. As in the case of the first formula the inappropriateness is greatest for small *N*'s and decreases as *N* increases.

Similar tests were made for the formula

$$\sigma_{r'} = \frac{1}{\sqrt{N-1}}. \tag{4}$$

The results are not presented here, but they indicate that this formula also yields too conservative an estimate of the significance of rhos for *N*'s up to 30.

It appears that although one or more of the above formulas might be useful when *N* is large, none of them is satisfactory for determining levels of significance of rhos obtained with *N*'s less than 30.

One method of testing the significance of rhos at the .01 and .05 levels has been found, however, which yields results surprisingly close to the empirical and approximated-empirical results from Olds' data. This method consists of using the values given by Guilford [2, pp. 323 f., adapted from Wallace and Snedecor (6), who in turn adapted their table from Fisher] for coefficients of correlation that are significant at .01 and .05 levels for various degrees of freedom and multiplying each value by the constant 1.04. The close agreement between the figures obtained by this method and the figures obtained from Olds' data may be seen in Tables 2 and 3. The only serious discrepancies for the .01 level of significance occur for an *N* of five and for *N*'s from 10 to about 15 or 18. In the case of these latter *N*'s we have already noted that there is reason to doubt the values obtained from Olds' data.

The discrepancies appear to be greater for the .05 level of significance, although still relatively small. Here we can compare the values obtained by taking 1.04 times Guilford's figures with values from Olds' data only for *N*'s up to 10, since Olds does not give the necessary data for *N*'s of 11 to 30. It is possible, however, to make rough interpolations from the data in Table 1 and obtain approximations for the coefficients that would be significant at the .05 level for *N*'s of 11 to 30. It appears that the discrepancies at the .05 level between the coefficients indicated by Olds' data and the values obtained by taking 1.04 times Guilford's figures increases as *N* increases from

TABLE 2

Results Obtained by Various Methods of Determining the Rhos That are Barely Significant at the .01 Level

N	2	3	4	5	6	7
2	none	......	......	*	.952†	none
3	none	.990	none	*	1.348†	none
4	none	.920	none	*	1.650†	none
5	none	.850	none	*	1.905†	.997
6	.960	.795	none	25.946	2.043	.954
7	.906	.752	none	11.766	2.112	.909
8	.869	.716	none	8.959	2.189	.867
9	.825	.687	none	6.933	2.224	.830
10	.788	.653	none	5.925	2.251	.796
11	.816		none	7.351	2.458	.764
12	.777		none	6.216	2.451	.736
13	.745		.941	5.519	2.459	.711
14	.720		.889	5.143	2.474	.687
15	.689	.570	.846	4.687	2.452	.667
16	.666		.807	4.411	2.458	.648
17	.645		.775	4.216	2.452	.630
18	.626		.745	4.039	2.455	.614
19	.608		.716	3.897	2.462	.598
20	.592	.512	.694	3.795	2.456	.583
21	.577		.672	3.675	2.455	.571
22	.563		.652	3.609	2.459	.558
23	.550		.634	3.526	2.455	.547
24	.538		.617	3.449	2.457	.536
25	.527	.469	.601	3.400	2.463	.525
26	.516	.462	.587	3.351	2.457	.516
27	.506	.455	.574	3.307	2.456	.506
28	.497	.449	.561	3.270	2.460	.497
29	.488	.442	.549	3.232	2.465	.489
30	.479	.436	.539	3.193	2.456	.482

* *t* ratio is indeterminately high.

† Calculated by assuming rho to be 1.00.

Column 1 lists the *N*'s.

Column 2 lists the rhos calculated from Old's empirical and approximated-empirical data (taken from Table 1 of this report).

Columns 3 and 4 list the rhos that would be significant according to Guilford's table of *t* ratios for *N*—2 degrees of freedom when the standard error of rho is assumed to be $\frac{1.05(1-r'^2)}{\sqrt{N-1}}$ (column 3) or $\frac{1.05}{\sqrt{N-1}}$ (column 4).

Columns 5 and 6 list the *t* ratios obtained by dividing the values of rho obtained from Old's data (given in column 2) by the standard error of rho found by the formula $\frac{1.05(1-r'^2)}{\sqrt{N-1}}$ (column 5) or $\frac{1.05}{\sqrt{N-1}}$ (column 6).

Column 7 gives the values obtained for significant rhos by taking 1.04 times the values given by Guilford for *r*'s that are significant at the .01 level (in Table 3, at the .05 level).

TABLE 3

Results Obtained by Various Methods of Determining the Rhos that are Barely Significant at the .05 Level. (The descriptions of the columns are the same as in Table 2 above.)

N	2	3	4	5	6	7
2	none	......	......	*	.952†	....
3	none	.948	none	*	1.347†	1.037
4	none	.827	none	*	1.650†	.988
5	.933	.745	none	13.721	1.777	.913
6	.843	.687	none	6.203	1.795	.843
7	.781	.644	none	4.671	1.822	.784
8	.725	.610	.971	3.851	1.827	.735
9	.681	.581	.878	3.420	1.835	.693
10	.644	.557	.807	3.145	1.840	.657

* Indeterminately high.

† Calculated by assuming rho to be 1.00.

7 up to 30. Even at *N* equals 30, however, the discrepancy appears to be less than .02. In the absence of a better test of significance, even at the .05 level, the values obtained by multiplying 1.04 times Guilford's figures might serve for N's up to 30. These values apparently are slightly too conservative from *N* equals 7 up, and become increasingly so as *N* increases. Despite this weakness, which should be noted, the values obtained in this manner from Guilford's figures have been listed in Table 1 as the least significant coefficients at the .05 level for *N*'s of 11 to 30.

It was noted earlier that the figures in Table 1 for the .01 level of significance are out of line for *N*'s of 11 ff., because of the distortion resulting from the assumption of normal distributions. The fortunate circumstance that the values obtained by 1.04 times Guilford's figures fit so closely the values in Table 1 for *N*'s up to 10 and for *N*'s of 26 to 30 suggests a simple method of obtaining an approximate correction for the distortion. This consists of substituting for the figures in the .01 column of Table 1 for *N*'s of 11 to 25, inclusive, the figures in Table 2 obtained by taking 1.04 times Guilford's values for coefficients of correlation that are barely significant at the .01 level for the corresponding *N*'s. This amounts to making rough interpolations of the values for *N*'s of 11 to 25 from the values for *N*'s up to 10 and from *N*'s of 26 to 30. The values obtained by means of this rough interpolation are listed in parentheses in Table 1.

Since values equal to 1.04 times Guilford's values agree so closely with the coefficients obtained from Olds' data for the .01 level of significance for *N*'s of 25 to 30, it would appear reasonable to assume that for N's immediately above 30 one can obtain a close approximation to the coefficients that are significant at the .01 level by taking 1.04 times Guilford's figures (2, pp. 323 f., after Wallace and Snedecor, who interpolated from Fisher's data).*

The Problem of Tie Rankings and Tests of Significance

Neither Olds (5) nor Kendall and others (4) consider the problem of tie rankings in relation to tests of significance for coefficients of rank difference correlation. Perhaps the problem is unimportant, and yet a casual survey of some of the combinations possible when tie rankings occur with an N of four suggests that when N is very small one or more pairs of tie rankings will change very greatly the frequencies with which various values of $\sum d^2$ and rho can be obtained. When N equals four, for example, there are at least seven ways of getting a $\sum d^2$ of zero (or a rho of + 1.00) when tie rankings occur. In addition there are at least seven ways of getting $\sum d^2$ greater than zero but less than two. By contrast, out of the 24 possible combinations obtainable with an N of four in data in which no tie rankings occur, there is only one way of obtaining a $\sum d^2$ of zero and three ways of obtaining $\sum d^2$ of two.

It seems probable that as N increases, and surely as the ratio of tie rankings to untied rankings decreases, the effect of tie rankings upon the probabilities of obtaining various values of $\sum d^2$ and rho will decrease in importance. It would be futile, however, to tabulate or calculate all the possible combinations that could occur for various N's in data involving tie rankings; for there is no possible way of determining adequately the probability of the ties themselves occurring.** The probability of ties occurring will vary with the type of

* Inspection of Tables 2 and 3 suggests the hypothesis that as N increases the formula 1.04 times Guilford's figures yields increasingly too conservative estimates of the coefficients that are significant at both the .01 and .05 levels. This tendency appears to begin at N equals 7 for the .05 level. Where it begins for the .01 level of significance is not clear from these data; perhaps it is at N equals 7, N equals 9, or N equals 29. It is possible that for large N's (probably larger for the .01 level than for the .05 level of significance) Guilford's figures for significant coefficients of correlation would apply directly without correction. The evidence for this suggestion is admittedly weak, although rough interpolation in Table 1 for the value for N of 30 at the .05 level of significance indicates that the value obtained would be approximately that given in Guilford's table.

** It would be possible to tabulate separately all of the possible combinations of pairs of ranks for various N's, first for data involving a single instance of a tie ranking involving two cases and occurring in only one of the two sets of ranks, second for data involving a single instance of a tie involving two cases in both sets of ranks, third for data involving a single set of ties between three cases in one set of ranks, and so on for all possible types, numbers, and combinations of ties. Tables of probabilities might then be set up separately for each of the numerous types of data involving ties. The task would be well nigh endless and seems scarcely worth while or practicable.

data involved and with the methods used in collecting and treating the data—whether tie judgments are permitted, to how many decimal places calculations are carried, etc.

Another problem arises in data involving tie rankings. For the case where N is four, there are at least seven ways of obtaining a rho of +1.00, but there is no way of obtaining a rho of —1.00 so long as ties are present in either set of ranks and the usual formula for rho is used. The question arises as to whether tie rankings tend to increase the probability of positive coefficients and to decrease the probability of negative coefficients. If so, it then becomes pertinent to inquire to what extent the probabilities of positive and negative coefficients are changed by the presence of ties. Perhaps it will be found necessary to introduce a change in the denominator of the formula for rho for data involving tie rankings.

It appears probable that the importance of the questions raised in the preceding paragraph will decrease as N increases and as the ratio of tie rankings to untied rankings decreases. Nonetheless, these questions along with the, for this paper, more pertinent general problems stated above should serve to create doubt concerning the adequacy of any method of testing the significance of rhos obtained from data in which tie rankings occur. In the absence of evidence concerning the probabilities of various values of rho for data involving ties, the levels of significance given in Table 1 probably represent the best means available at present for testing the significance of such rhos. When N is relatively large and the number of ties is relatively small, the problem of the effect of ties upon the significance of a coefficient can probably safely be ignored. When N is small and when the number of ties is relatively large, it perhaps should be obligatory upon an experimenter to report the number and nature of the tie rankings as well as the rho and the N. Certainly in this latter case any statement concerning the level of significance of a rho is open to question.

Recommendations

From the data presented in this paper a few recommendations concerning tests of significance of rank difference coefficients of correlation seem justified:

(1) When calculations are made by the usual procedure using Spearman's formula, Old's (5, pp. 145-148) tables of probabilities for various values of $\sum d^2$ for N's from 2 to 30 are the most convenient and adequate means of testing significance. The few minor errors in Olds' tables, noted earlier in this paper, should be corrected, however; and the distortions present in the probabilities for N of 11 and N's

immediately above 11 should be noted.

(2) When testing the significance of rhos reported in the literature or calculated by methods not involving the finding of $\sum d^2$, the values given in Table 1 of this report will be useful for N's from 2 to to 30. When testing significance at the .01 level, where alternate values are given in Table 1, the values given in parentheses are preferable since they have been corrected for the distortion present in Olds' probabilities for N's of 11 ff. In testing significance at other levels than .01 for N's of 11 and above, the levels of .04 and .10 are preferable because the values appear to be more accurate for these levels of significance than for the .02 and .05 levels. For N's of 10 and below the values in Table 1 may be considered quite exact—completely so for N's of 8 and below for coefficients obtained from data not involving tie rankings.

(3) A close approximation to the rank difference coefficients that are barely significant at the .01 level for N's immediately above 30 is probably obtained by taking 1.04 times the values given for $N-2$ degrees of freedom in Guilford's adaptation (2, pp. 323 f.) of Wallace and Snedecor's table of coefficients of correlation that are significant at the .01 level. Perhaps the values given in the same table for coefficients significant at the .05 level may be taken directly as the rhos that are significant at the .05 level for N's immediately above 30; this latter is less well justified by the evidence, however.

(4) The formulas $\dfrac{1.05(1 - r'^2)}{\sqrt{N-1}}$, $\dfrac{1.05}{\sqrt{N-1}}$, and $\dfrac{1.00}{\sqrt{N-1}}$ for calculating a standard error of a rho appear inaedquate for purposes of calculating t ratios by which to test the significance of a rank difference coefficient of correlation for N's up to 30. The first formula results in too high t ratios; the other two formulas result in t ratios that are too low.

(5) When tie rankings occur in the data from which a rho is calculated, caution should be used in evaluating the level of significance of the coefficient. The data given in Table 1 of this report probably represent the best means available at present for testing the significance of such a coefficient; but whenever the number of tie rankings is relatively large in comparison to the N of the ranks, and especially when N is very small, any test of significance should probably be considered questionable.

REFERENCES

1. DuBois, P. H. Formulas and tables for rank correlation. *Psychol. Rec.*, 1939, 3, 46-56.
2. Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1942.
3. Hotelling, H. and Pabst, M. Rank correlation and tests of significance involving no assumption of normality. *Ann. math. Statist.*, 1936, 7, 29-43.
4. Kendall, M. G., Kendall, S. F. H., and Smith, B. B. The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, 1938, **30**, 251-273.
5. Olds, E. G. Distributions of sums of squares of rank differences for small numbers of individuals. *Ann. math. Statist.*, 1938, 9, 133-148.
6. Wallace, H. A. and Snedecor, G. W. *Correlation and machine calculation.* Ames, Iowa: Iowa State College, 1931.

PSYCHOMETRIKA—VOL. 8, NO. 4
DECEMBER, 1943

A COURSE IN THE THEORY OF MENTAL TESTS

HAROLD GULLIKSEN*
PSYCHOLOGY DEPARTMENT
THE UNIVERSITY OF CHICAGO

An outline for a course in test theory is presented, together with a list of assignments, problems, and a bibliography. The course has been given in the Psychology Department of the University of Chicago. The material is presented in outline form at the present time because of the increased need for training in test theory due to the increase in the use of psychological tests for classification of military personnel, and because much of the material in such a course must be selected from a wide array of articles in the literature. This material is presented in order that an organized body of material for instructional purposes may be readily available to those interested.

The recent increase in psychological testing for the purpose of classifying military personnel has increased the need for training in the quantitative theory which is basic to the construction of tests and the analysis of test data.

During the last forty years, progress has been made toward a well-integrated, quantitative theory pertaining to the behavior of test items and test scores under different conditions. This rationale of mental tests should not be confused with "statistics." A good foundation in elementary statistics is a *prerequisite* for a course in the theory of mental tests.

At present the work on the theory of mental tests is for the most part not available in text-book form. It is necessary to select the reading material for such a course from a wide array of journal articles. The writer has prepared an outline, bibliography, and problems for such a course. This material is presented here in order that an organized body of material on test theory for instructional purposes may be readily available to those interested.

This course deals with the mathematical and statistical theory necessary for interpreting test results. It does not deal with the non-quantitative problems involved in the construction of aptitude or achievement tests; nor does it attempt to familiarize the student with the various psychological and educational tests now available. How-

* On leave from the University of Chicago for a government research project at the College Entrance Examination Board, Princeton, New Jersey.

ever, reference is made to articles and books dealing with material of this nature, and the student is urged to familiarize himself with this material in case he has not already had work dealing with these topics.

The material on test theory is presented in four sections. First an outline of the course, then a specific list of assignments, followed by a set of problems and the bibliography. In the preparation of the bibliography, a good foundation in elementary statistics, including correlation, was assumed. For some of the articles a knowledge of multiple correlation and matrix theory is necessary. It was found desirable in preparing the material for class use to have the assignments as compactly presented as possible, so that the student could easily see just what had to be done. The assignments follow the same numbering system as the outline, so that it is possible to refer readily to the assignment for any part of the outline, or to look up that part of the outline corresponding to any assignment.

Outline

I. Introduction to the course and review of statistics.

(six class periods, with review examinations on statistics in the third and fifth periods. The examinations are returned and discussed during the fourth and sixth sessions.)

II. Accuracy of test scores.

A. Different measures of accuracy of test scores.

1. Standard error of measurement.

a. Estimation of test error usually made by using two comparable tests on the same population.

b. Assumptions usually used to define test error—the basic assumptions of test theory.

(chance error distinguished from constant and systematic errors)

(1) The obtained score s of any person may be divided into two parts; a true score t which represents his actual ability, and an error score e which is due to the various factors that may cause a person to answer correctly an item which he does not know, or to answer incorrectly an item which he does know. Specifically it is assumed that these three scores, s, t and e are related as follows:

$$s = t + e, \quad \text{or} \quad e = s - t.$$

(2) The sum of chance error scores for the group, and hence the average error, is zero. This may be written

$$\sum e = 0\,, \quad \text{or} \quad M_e = 0\,.$$

(3) The correlation of error scores with true scores is zero. This may be written

$$\sum et = 0\,, \quad \text{or} \quad r_{et} = 0\,.$$

(4) The correlation of error score for one test with error score for another test is zero. A special case of this assumption is that the correlation between error scores on two parallel forms of the same test is zero. This may be written

$$\sum e_1 e_2 = 0\,, \quad \text{or} \quad r_{e_1 e_2} = 0.$$

(5) The standard deviation of the distribution of error scores on one form of a test is equal to the standard deviation of the distribution of error scores on a parallel form of that test. This may be written

$$\sum e_1^2 = \sum e_2^2, \quad \text{or} \quad \sigma_{e_1} = \sigma_{e_2}.$$

(6) The true score on one test for a given person is equal to the true score of that person on a parallel form of the test. That is

$$t_1 = t_2\,.$$

[From assumption (6) it follows that

$$\sigma_{t_1} = \sigma_{t_2}, \quad M_{t_1} = M_{t_2}, \quad \text{and}$$
$$r_{t_1 t_2} = 1.]$$

c. Estimation of parameters of distribution of true scores, and of error scores.

d. Definition of equivalent forms of a test.

(1) Items in form 1 psychologically similar to items in form 2.

(2) $M_1 = M_2\,.$

(3) $\sigma_1 = \sigma_2\,.$

e. Test variance equals sum of true variance and error variance.

f. Interpretation of standard error of measurement.

(1) Standard deviation of distribution of error scores.

(2) Error made in using obtained scores instead of true scores.

(3) Error made in predicting obtained scores from true scores.

(4) Obtained variance minus true variance.

2. Reliability coefficient and index of reliability.
   a. Standard error of measurement expressed as a function of test variance and reliability.
   b. Reliability coefficient as ratio of true to total variance.
   c. Interpretation of coefficient and index of reliability.

3. Standard error of estimate.
   a. Of test score.
   b. Of true score.

4. Standard error of substitution.
   a. Derivation.
   b. Interpretation.

5. Coefficient of alienation.

6. Comparison of various measures of accuracy of test scores.

B. Factors affecting accuracy of test scores.

1. Effect of increased dispersion on reliability and standard error of measurement.
   a. Increased dispersion due to increased range of ability.
   b. Increased dispersion due to increased test error.

2. Length of test.
   a. Effect of doubling test length.
   b. Effect of further increasing length of test.
   c. Prediction of length necessary for any given reliability.
   d. Experimental verification of *c*.

3. Difficulty of test.

4. Inter-item *r*'s.

C. Methods of estimating test reliability.

1. Repetition of same form.
2. Equivalent forms.
3. Split-halves method.
4. Reliability of time-limit tests.
5. Method for unique estimation of reliability from parameters of test score distribution.
6. Estimating reliability from mean, variance, and number of items.
7. Reliability coefficient calculated from variance of sums and variance of differences of halves.

D. Special topics in reliability.

1. Reliability of a sum of variables.
   r of any sum. Various special cases of this.
2. Reader reliability of a test.
3. Content reliability of a test.
4. Common factor interpretation of reliability.

E. Validity of a test.

1. Definition of validity (correlation of test with criterion).
2. Relation to the concept of reliability.
3. Relationship between length and validity.
4. Length required for given validity.
5. Relationship between reliability and validity.
6. Correction for attenuation.
7. Relation to difficulty.

III. Scoring Methods.

A. Linear transformations.
1. Deviation score.
2. Standard scores.
3. Derived scores with any given mean and standard deviation.

B. Non-linear transformations.

1. Percentile.
2. Normalized scores.
3. McCall's T-score.
4. Scaled scores of the Cooperative Test Service.

C. Age scales and other scales depending upon an external criterion.

1. Mental age.
2. Intelligence quotient.
3. Criticism of the mental age concept.

D. Absolute scaling.

1. Factors affecting the shape of the test score distribution.
    a. Number of items.
    b. Inter-item correlation.
    c. Item difficulty distribution.
    d. Test score is not necessarily linear with any true measure of ability.
2. Problem of a constant unit of measurement (difficulty and the discrimination function).
3. An absolute zero point.
4. Applications of absolute scaling methods.

E. Scoring formulas.

1. Correction for chance successes.
2. Empirical scoring formulas using multiple correlation.
3. Time limit vs. amount limit tests.
    a. Power vs. speed tests.
    b. Variance of distribution of number omitted.
4. Scoring of rank-order items.

IV. Combination of Measures.

A. Combination of test scores.

1. To maximize validity, given an external criterion.
    a. Multiple correlation.
       Caution on multiple correlation.
    b. Approximations to multiple correlation (see IV, B, 1, b).
    c. Group difference and other methods (see IV, B, 1, c).
2. To maximize reliability or internal consistency when no external criterion is available.
    a. Inter-test correlations.
       Maximizing inter-individual variance.
       Minimizing inter-test variance.
       The one-factor approach.
       Minimizing generalized variance of all individuals receiving the same score.

b. Use of approximations to inter-test correlations (see IV, B, 2, b).

c. Other methods.
By dispersions.
By reliabilities.
By difficulty.

3. Irrelevance of weights when number of variables is large or when inter-correlations are high.

4. Comparison and criticism of methods (see IV, B, 4).

B. Selection of test items.

1. To maximize validity (given an external criterion).

a. Multiple correlation (see IV, A, 1, a)
(usually too laborious when dealing with items).

b. Approximations to multiple correlation.
L-method.
Successive residuals.
Maximizing function.
Rapid approximation to L-method.
Flanagan's method.

c. Group difference and other methods.
Weighting of items in a
personality test.
interest test.
application blank.
(weighting in a test where best ordering of test responses cannot be prejudged).
Fisher, analysis of variance.
With three or more possible answers weight to maximize correlation.
Theoretically the multiple correlation approach is best, but in many cases it is too laborious. Other methods may be regarded as an approximation to this method, either to save work, or because the criterion is missing.

2. To maximize reliability or internal consistency when no external criterion is available.

a. Use of inter-item correlations (see IV, A, 2, a).
Usually too laborious when dealing with items.

b. Use of approximations to inter-item correlations.
Use of item-test correlation.

c. Other methods (see IV, A, 2, c).

3. Irrelevance of weights when number of variables is large, or when inter-correlations are high (see IV, A, 3).

4. Comparison and criticism of methods.
Guilford.
Adkins.
Long.
Lentz.
Merrill (homogeneity).
Mosier (one-factor approach).

V. Relationship between test theory and psychophysics.

Assignments

General directions for assignments

Textbook and references.

The principal text is Thurstone, 1931, *The Reliability and Validity of Tests.* This book is referred to as Thurstone 1931. J. P. Guilford's *Psychometric Methods* is recommended for those who wish to use two texts. The references in the assignments are given by author and date. The complete reference may be found in the bibliography, listed alphabetically by author's name.

Problems.

The problems accompanying the assignments are either from Thurstone's text, or from the list of supplementary problems given after the assignments. The former are indicated by numbers, the latter by letters.

Suggestions for preparing the problems

1. The following materials will be needed: ruler, compass, graph paper, data sheet paper, (11" x 17").
2. Leave a 1 1/2 inch margin so that the papers can be bound without concealing the work.
3. Make all graphs on printed graph paper.
4. The best way to bind the data sheets is to fold them with the ruled surface *out* and put both edges into the binding. This necessitates leaving a 1 1/2-inch margin on both the right and left edges and not using the column which comes at the fold.
5. In making graphs use an x for each point. This method makes it easy to see the general trend of the graph and yet secures accuracy since the intersection of the two lines of the cross is placed at the exact point to be indicated.

6. Make all computations neatly on the *first copy* so that the original data sheet can be handed in. Never do your figuring on scratch paper and then copy it to hand in. This method both takes additional time and increases the chances of error.
7. Label *each* problem, giving the *book, chapter, page,* and *problem number.*
8. The problems for each assignment should be bound in a folder and handed in as a unit with your *name* and the assignment indicated.

Assignment I

Readings: Thurstone 1924 Ch. 1-3, 6, 7, 10-15, 17-19; Ch. 4, 8, 9, 21-25

Assignment II-A

Readings: Thurstone 1931 preface, pp. 1-4, 17-20
Boring, 1920

Topic 2

Readings: Thurstone 1931 pp. 12-16, 23
Guilford 1936 (b) pp. 408-421
Problems: Thurstone 1931 (8, 9, 10, 11, 12, 18, 25, 26)

Topic 3

Readings: Thurstone 1931 pp. 52-56
Douglass 1934
Monroe 1934
Problems: Exercise B

Topic 4

Readings: Kelley 1927 pp. 211-213, 171-181

Topic 5

Readings: Thurstone 1931 pp. 57-61
Problems: Thurstone 1931 (2, 16)

Topic 6

No special readings or problems

Assignment II-B

Topic 1

Readings: Thurstone 1931 pp. 24-27
Otis 1922 (b)
Problems: Thurstone 1931 (28, 29, 30, 31)

Topic 2

Readings: Thurstone 1931 pp. 28-36, 40-45
Holzinger and Clayton 1925
Ruch et. al. 1926
Problems: Thurstone 1931 (5, 6, 7, 13, 17, 20, 22, 34, 35, 36)

Topic 3

No special readings or problems

Topic 4

Readings: Richardson 1936 (a)

Assignment II-C
Topic 1-4
Readings: Thurstone 1931 pp. 5-11

Topics 5-6
Readings: Kuder and Richardson 1937
Dressel 1940
Richardson and Kuder 1939
Problems: Exercise K

Topic 7
Readings: Rulon 1939
Otis and Knollin 1921
Problems: Exercise H

Assignment II-D
Topic 1
Readings: Kelley 1924 pp. 196-200
Walker 1929 pp. 112-118

Topic 2
Readings: Starch and Elliot 1912, 1913 (a), 1913 (b)
Monroe 1938

Topic 3
Readings: Gulliksen 1936
Problems: Exercise J

Topic 4
Readings: Burt, Sections I-V, emphasizing pp. 280-297

Assignment II-E
Topics 1-2
Readings: Thurstone 1931 pp. 97-104
Otis 1922 (a)
Guilford 1936 (b) pp. 421-426
Segel 1933
Problems: Thurstone 1931 (3, 4, 24)

Topics 3-4
Readings: Thurstone 1931 pp. 46-49
Problems: Thurstone 1931 (21, 37, 38, 39)

Topic 5
Readings: Thurstone 1931 pp 50-51
Problems: Thurstone 1931 (14, 33, 40, 41)

Topic 6
Readings: Thurstone 1931 pp. 62-68
Walker 1929 pp. 118-123
Spearman 1907, 1910
Brown and Thomson 1921 Ch. 8
Problems: Thurstone 1931 (1, 15, 23, 42, 43, 47)
Exercise A

Topic 7
Readings: Richardson 1936 (b)
Thurstone, T. G. 1932

- Assignment III-A
  - Topics 1-3 incl.
    - Readings: Thurstone 1924 Ch. 15
    - Problems: Exercise C
- Assignment III-B
  - Topics 1-2
    - Readings: Thurstone 1924 Chs. 16 and 20
    - Problems: Exercise D
  - Topics 3 and 4
    - Readings: McCall 1922 Ch. 10
      - Flanagan 1939
- Assignment III-C
  - Topics 1-3
    - Readings: Freeman 1917
      - Freeman 1939 Ch. IV
      - Thurstone 1926
      - Toops and Symonds 1922
      - Yerkes, et. al. 1915
      - Stern 1914 p. 80
- Assignment III-D
  - Topics 1-2
    - Readings: Kelley 1923
      - Thurstone 1925
    - Problems: Exercise E
  - Topic 3
    - Readings: Thurstone 1928
  - Topic 4
    - Readings: Flanagan 1939
      - Thorndike 1922
- Assignment III-E
  - Topics 1-4
    - Readings: Thurstone 1931 pp. 69-82
      - Thurstone 1919
      - Moore 1940
      - Guilford 1936 (a)
    - Problems: Thurstone 1931 (19)
      - Exercises F and I
- Assignment IV-A
  - Topics 1
    - Readings: Garrett 1943
      - Guilford 1936 (b) pp. 380-399
      - Frisch 1934
    - Problems: Exercise G
  - Topic 2 (a)
    - Readings: Burt 1936 (Section VI, pp. 297-304)
      - Edgerton and Kolbe 1936
      - Horst 1936 (a)
      - Wilks 1938

Topic 2 (c)
Readings: Thurstone 1931 pp. 83-90
Richardson: 1936 (b)
Thurstone, T. G. 1932
Problems: Thurstone 1931 (44, 45)

Topic 3
Readings: Stalnaker 1938
Stalnaker and Richardson 1933
Wilks 1938
Burt 1936
Lee and Symonds 1934 (for review of studies on weighting)

Assignment IV-B
Topic 1 (b)
Readings: Adkins and Toops 1937
Flanagan 1936
Horst 1934 (a); 1934 (b); 1936 (b)

Topic 1 (c)
Readings: Thurstone 1931 pp. 91-96
Boring 1919
Lindquist 1940
Travers 1939
Problems: Thurstone 1931 (32, 46, 48, 49)

Topic 2 (b)
Readings: Babitz and Keys 1940
Edgerton and Toops 1928
Richardson 1936 (a)

Topic 4
Readings: Guilford 1936 (b) pp. 426-437
Adkins (Thesis)
Lentz, et al 1932
Long, et al 1935
Merrill 1937
Mosier 1936
Swineford 1936

Assignment V
Readings: Guilford 1937
Mosier 1940

For bibliographies on tests and testing see
Buros 1936, 1937, 1939, 1941
Hildreth 1939
Holmes 1917
Lee and Symonds 1934
National Society for Study of Education 1918
Ruger 1918
Whipple 1910

For methods of constructing tests and test items see
Board of Examinations Chicago 1937
Englehart 1942

Hawkes, et al 1936
Hull 1928
Orleans 1937

Notes on arithmetic and graphic computing methods
Dunlap and Kurtz 1932
Dunlap 1936 (a) and (b)
Kuder 1937
Richardson 1935

General discussion of tests
Freeman 1939
Kelley 1927
McCall 1922
Monroe 1923
Monroe and Englehart 1936
Orleans 1937
Ruch and Stoddard 1927
Smith 1938
Stern 1914

PROBLEMS

Problem A

Derive the equation for predicting scores in X_1 from scores in X_2 where X_1 and X_2 are two different forms of a test whose reliability and validity are both known. Derive the equation for predicting the true score Y from Y_1, where Y_1 is a test whose reliability and validity are known.

Problem B

Describe the correct experimental method for checking the applicability of the formula

$$\sigma_e = \sigma_y \sqrt{1 - r_{xy}^2}$$

where σ_e is the standard error of estimate made in estimating y from x;
σ_y is the standard deviation of the observed y distribution; and
r_{xy} is the correlation between x and y.

PROBLEM C

Compute the table and draw the graph which would be used for transforming raw scores in the A. C. E. (1937 edition) (see information in accompanying table) into:

a. deviation scores (*d*-scores)
b. standard scores (*z*-scores)
c. derived scores with a mean of 50 and a standard deviation of 10 (*D*-scores)

Consider *only* the *total* distribution given in the right-hand column of the table.

TABLE FOR USE IN CONNECTION WITH PROBLEMS C AND D

The following scores were made by 68,899 students in 323 colleges on the 1937 edition of the American Council on Education Psychological Examination for College Freshmen:

		Frequency	
Scores	*Men*	*Women*	*Total**
0-9		2	4
10-19	5	3	12
20-29	27	22	58
30-39	85	50	170
40-49	169	112	329
50-59	329	225	626
60-69	471	358	943
70-79	667	479	1,314
80-89	923	769	1,915
90-99	1,108	892	2,264
100-109	1,387	1,171	2,897
110-119	1,669	1,376	3,429
120-129	1,768	1,529	3,764
130-139	2,064	1,768	4,348
140-149	2,113	1,793	4,471
150-159	2,188	1,830	4,650
160-169	2,220	1,748	4,600
170-179	2,128	1,798	4,583
180-189	1,990	1,610	4,207
190-199	1,823	1,479	3,904
200-209	1,639	1,351	3,593
210-219	1,488	1,251	3,281
220-229	1,234	996	2,686
230-239	1,097	906	2,441
240-249	893	748	2,025
250-259	750	596	1,630
260-269	584	488	1,309
270-279	474	329	998
280-289	358	273	772
290-299	284	187	580
300-309	184	122	387
310-319	153	74	286
320-329	96	52	181
330-339	70	38	133
340-349	29	13	51
350-359	24	9	40
360-369	6	3	14
370-379	2		2
380-389	1		2
Total	32,500	26,450	68,899
Lower quartile	127.27	127.54	128.67
Median	165.75	164.84	167.08
Upper quartile	207.57	206.10	208.87

* The total includes the scores of 9,949 students not classified according to sex. Data taken from Thurstone, L. L. and Thurstone, T. G. The 1937 Psychological Examination for College Freshmen. The Educational Record, April, 1938, pp. 209-234.

PROBLEM D

Compute the table and give the graph which would be used to change raw scores on the A. C. E. (1937 edition) into

a. percentile scores (p-scores)
b. normalized scores (N-scores)

Again use *only* the *total* distribution given in the right-hand column. Referring also to problem C draw the graphs showing the relation between

a. *D*-scores and *z*-scores
b. *D*-scores and *p*-scores
c. *D*-scores and *N*-scores
d. *z*-scores and *p*-scores
e. *z*-scores and *N*-scores
f. *p*-scores and *N*-scores
g. Plot *p*-scores against *D*-scores on arithmetic probability paper as a test of normality

Write a brief paragraph stating the relationship shown in the foregoing six graphs.

PROBLEM E

Below is given the frequency distribution of A. C. E. scores for 113 students taking the 56 tests used in Mr. Thurstone's first large study of primary mental abilities. (Thurstone, L. L. Primary mental abilities, p. 19). This distribution is given in terms of percentile points on the national norms for the A. C. E. test. The table shows that there was one student between the 35 and 40 percentile points on the national norms; two students between the 45 and 50 percentile points; and so forth. It will be noticed that over 25 per cent of the students are above the 98 percentile point on the national norms. Can this distribution of 113 cases be regarded as a normal distribution, granted the assumption of a Gaussian distribution of intelligence in the forty thousand students on whom the national norms were based? Use the absolute scaling methods to answer this question.

*scores**	*frequency*	*cumulative frequency*
35-40	1	1
40-45	1	2
45-50	2	4
50-55	2	6
55-60	4	10
60-65	6	16
65-70	2	18
70-75	4	22
75-80	10	32
80-85	6	38
85-90	10	48
90-95	25	73
95-96	3	76
96-97	4	80
97-98	3	83
98-99	6	89
99-999	23	112
999-1.000	1	113

* National norms (1933), 40,229 cases, 203 colleges.

PROBLEM F

Derive the formula for the correlation between number correct and number incorrect for an objective test, assuming that there are no omissions.

PROBLEM G

Entering freshmen at the University of Chicago are given an A .C. E. Psychological Examination (a), a physical sciences aptitude test (s), an English placement test (e). A year later they are given the physical science comprehensive (p), and the humanities comprehensive (h).

The following zero-order correlations are obtained:

$$r_{as} = .50, \quad r_{ae} = .70, \quad r_{es} = .40,$$
$$r_{ap} = .50, \quad r_{sp} = .70, \quad r_{ep} = .40,$$
$$r_{ah} = .60, \quad r_{sh} = .20, \quad r_{eh} = .70, \quad r_{ph} = .60.$$

The following means and standard deviations are found:

	a	s	e	p	h
Mean	120	110	150	220	460
Standard deviation	30	20	25	30	40

1. Write the equation for making the best prediction of the humanities comprehensive score from the three placement tests.
2. What will be the correlation between the predicted humanities scores and the actual scores, using the prediction formula given in 1?
3. Which two placement tests will give the best prediction of scores in the physical-science comprehensive?
4. Write the equation for making the best prediction of the physical-science comprehensive score from the two tests mentioned in 3.
5. What is the correlation between the actual physical-science scores and the scores predicted by using the equation given in 4?

PROBLEM H

Make the following calculations from the data given in the appendix.

Estimate the reliability coefficient for the total test by the split-half method using the Spearman-Brown correction.

Estimate the reliability coefficient from the variance of the total score and the variance of the difference between scores on the two halves.

Give the standard error of measurement for this test.

What are the probable limits for the true score of a person who scores 51; one who scores 73.

Estimate the reliability of a test which is twice as long as this one, four times as long, seven times as long, ten times as long (see table for Spearman-Brown formula in Dunlap and Kurtz). Estimate the reliability the test would have if it were made infinitely long. If one wished this test to have a reliability of .97, how long would it be necessary to make it? Graph the foregoing results.

Estimate the reliability this test would have if it were applied to a group whose scores had a standard deviation half that of the original group. Estimate the reliability this test would have if it were applied to a group whose scores had a standard deviation twice that of the original group.

PROBLEM I

Comment briefly on the material in Moore's 1940 article.

PROBLEM J

An examination has an objective section (o) and an essay section (e). The split-half correlation of scores for section o is .80. The corresponding correlation for section *e* is .65. On section *e* the correlation between the total score given by reader A and reader B is .85.

Estimate the content reliability for section *o*; for section *e*.

PROBLEM K

From the data given in the appendix, estimate the reliability of the test by the Kuder and Richardson method. Use equations 8, 14, 20, and 21. (Kuder and Richardson 1937).

APPENDIX

The following data on 52 students taking the composition section of the French 104-5-6 in June, 1940 were made available by Dr. Lawrence Andrus of the Board of Examinations at the University of Chicago.

Column A gives the item number.
Column B gives the proportion passing above mean.
Column C gives the proportion passing below mean.
Column D gives the proportion of entire group passing.
Column E gives r (tetrachoric).

A	B	C	D	E	A	B	C	D	E
1	.21	.04	.13	.51	51	.07	.00	.04	....
2	.68	.38	.54	.46	52	.79	.58	.69	.37
3	.57	.25	.42	.49	53	.46	.42	.44	.06
4	.68	.42	.56	.40	54	.89	.67	.79	.47
5	.82	.63	.73	.36	55	.39	.08	.25	.59
6	.32	.21	.27	.21	56	.29	.08	.19	.48
7	.71	.58	.65	.22	57	.68	.33	.52	.52
8	1.00	1.00	1.00	.00	58	.71	.29	.52	.61
9	.39	.08	.25	.59	59	.25	.13	.19	.09
10	.29	.13	.21	.35	60	.93	.58	.77	.70
11	.64	.42	.54	.34	61	.68	.50	.60	.29
12	.57	.17	.38	.62	62	.64	.42	.54	.34
13	.75	.42	.60	.51	63	.75	.38	.58	.56
14	.68	.50	.60	.29	64	.89	.96	.92	....
15	.93	.58	.77	.70	65	.75	.33	.56	.62
16	.79	.42	.62	.58	66	.57	.25	.42	.49
17	.64	.54	.60	.16	67	.50	.25	.38	.41
18	.75	.46	.62	.46	68	.50	.17	.35	.54
19	.96	.79	.88	.54	69	.46	.42	.44	.06
20	.57	.25	.42	.49	70	.36	.17	.27	.36
21	.86	.79	.83	.17	71	.75	.33	.56	.62
22	.64	.58	.62	.09	72	.54	.50	.52	.06
23	.75	.42	.60	.51	73	.75	.33	.56	.62
24	.86	.67	.77	.39	74	.36	.08	.23	.57
25	.71	.54	.63	.28	75	.71	.58	.65	.22
26	.64	.17	.42	.69	76	.54	.58	.56	–.06
27	.79	.29	.56	.72	77	.29	.21	.25	.15
28	.86	.79	.83	.17	78	.54	.21	.38	.52
29	.57	.42	.50	.23	79	.32	.17	.25	.29
30	.86	.71	.79	.33	80	.61	.25	.44	.54
31	.71	.67	.69	.07	81	.86	.63	.75	.45
32	.36	.17	.27	.36	82	.39	.17	.29	.40
33	.68	.33	.52	.52	83	.43	.21	.33	.38
34	.68	.54	.62	.23	84	.54	.17	.37	.58
35	.75	.75	.75	.00	85	.89	.50	.71	.68
36	.89	.71	.81	.42	86	.61	.63	.62	–.03
37	.64	.88	.75	–.47	87	.79	.58	.69	.37
38	.43	.46	.44	–.05	88	.39	.08	.25	.59
39	.64	.21	.44	.64	89	.82	.50	.67	.53
40	1.00	1.00	1.00	.00	90	.61	.54	.58	.11
41	.96	1.00	.98	....	91	.14	.00	.08	....
42	.71	.71	.71	0	92	1.00	.79	.90	.80
43	.71	.63	.67	.14	93	.79	.21	.52	.79
44	.82	.42	.63	.62	94	.93	.67	.81	.60
45	.82	.75	.79	.15	95	.86	.79	.83	.17
46	.79	.75	.77	.08	96	.96	.63	.81	.77
47	.64	.50	.58	.22	97	.71	.13	.44	.80
48	.57	.71	.63	–.23	98	.14	.00	.08	....
49	.14	.21	.17	–.18	99	.86	.71	.79	.33
50	.07	.04	.06	....	100	.79	.29	.56	.72

I	II	III	IV
1	41	24	17
2	40	22	18
3	73	40	33
4	39	20	19
5	74	37	37
6	49	31	18
7	35	20	15
8	59	33	26
9	44	28	16
10	51	25	26
11	55	26	29
12	54	31	23
13	36	25	11
14	74	35	39
15	48	29	19
16	52	28	24
17	66	42	24
18	73	39	34
19	59	33	26
20	50	26	24
21	25	18	7
22	60	31	29
23	60	34	26
24	65	34	31
25	41	18	23
26	65	35	30
27	58	32	26
28	35	24	11
29	55	31	24
30	62	32	30
31	68	32	36
32	55	30	25
33	62	29	33
34	67	36	31
35	53	30	23
36	54	29	25
37	61	32	29
38	68	31	37
39	58	30	28
40	60	29	31
41	84	43	41
42	39	20	19
43	37	22	15
44	56	28	28
45	56	31	25
46	25	15	10
47	72	36	36
48	33	24	9
49	41	26	15
50	66	35	31
51	38	27	11
52	84	41	43

Column I gives the code number of each student.
Column II gives the total score for each student.
Column III gives the score on the first fifty items for each student.
Column IV gives the score on the second fifty items for each student.

Number of items = 100.
Maximum number of score points = 100.
Mean raw score = 54.52.
Standard deviation = 13.98
Number of students = 52.

BIBLIOGRAPHY

1. Adkins, Dorothy C. A comparative study of methods of selecting items. Dissertation on file in library of Ohio State University. Abstract, Psychology Library.
2. Adkins, Dorothy C., and Toops, Herbert A., 1937. Simplified formulas for item selection and construction. *Psychometrika*, 2, 165-171.
3. Ayres, Leonard P., 1911. A scale for measuring the quality of handwriting of school children. New York: Publication on Measurement in Education, Division of Education, Russell Sage Fund Bulletin No. 113.
4. Babitz, Milton, and Keys, Noel. 1940. A method for approximating the average intercorrelation coefficient by correlating the parts with the sum of the parts. *Psychometrika*, 5, 283-288.
5. Board of Examinations, The University of Chicago. 1937. Manual of Examination Methods. Second Edition. Chicago: Univ. Chicago Bookstore. Pp. 177.
6. Boring, E. G. 1919. Mathematical vs. scientific significance. *Psychol. Bull.*, 16, 335-338.
7. ———. 1920. The logic of the normal law of error in mental measurement. *Amer. J. Psychol.*, 31, 1-33.
8. Bradford, Leland P. 1940. The effect of practice upon standard errors of estimate. *Psychol. Monogr.*, 52, No. 3, 56-71.
9. Brown, William, and Thomson, Godfrey. 1921. Essentials of mental measurement. Cambridge Univ. Press. Pp. viii + 216.
10. Buros, Oscar K. 1936. Educational, Psychological, and Personality Tests of 1933, 1934, and 1935. New Brunswick, New Jersey: School of Education, Rutgers University. Pp. 83. Reviews 1-503.
11. ———. 1937. Educational, Psychological, and Personality Tests of 1936. New Brunswick, New Jersey: School of Education, Rutgers University. Pp. 141. Reviews 504-868.
12. ———. 1938. The 1938 Mental Measurements Yearbook. New Brunswick, New Jersey: School of Education, Rutgers University. Pp. xiv + 415. Reviews 869-1181.
13. ———. 1941. The 1940 Mental Measurements Yearbook. New Brunswick, New Jersey: School of Education, Rutgers University. Pp. xxi + 674. Reviews 1182-1684.
14. Burt, Cyril. 1936. Supplement. In "The Marks of Examiners" by Hartog, P. J., and Rhodes, E. C. London: Macmillan and Company. Pp. xix + 344.
15. Douglass, H. R. 1934. Some observations and data on certain methods of measuring the predictive significance of the Pearson product-moment coefficient of correlation. *J. educ. Psychol.*, 25, 225-232.
16. Dressel, Paul L. 1940. Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 5, 305-310.
17. Dunlap, Jack W. 1936 (a). Note on the computation of bi-serial correlation in item evaluation. *Psychometrika*, 1, 51-58.
18. ———. 1936 (b). Nomograph for computing bi-serial correlations. *Psychometrika*, 1, 59-60.
19. Dunlap, Jack and Kurtz, A. K. 1932. Handbook of statistical nomographs and formulas. New York: World Book Company. vii + 163.
20. Edgerton, H. A. and Toops, H. A. 1928. A formula for finding the average inter-correlation coefficient for unranked raw scores without solving any of the individual intercorrelations. *J. educ. Psychol.*, 19, 131-138.

21. Edgerton, H. A. and Kolbe, Laverne E. 1936. The method of minimum variation for the combination of criteria. *Psychometrika*, 1, 183-187.

22. Englehart, Max D. 1942. Unique types of achievement test exercises. *Psychometrika*, 7, 103-115.

23. Flanagan, John C. 1936. A short method for selecting the best combination of test items for a particular purpose. *Psychol. Bull.*, 33, 603-604.

24. ———. 1939. Scaled scores. New York: Cooperative Test Service.

25. Freeman, Frank N. 1917. A critique of the Yerkes-Bridges-Hardwick comparison of the Binet-Simon and point scales. *Psychol. Rev.* 24, 484.

26. ———. 1939. Mental tests: Their history, principles, and applications. Cambridge, Mass.: The Riverside Press, Rev.

27. Frisch, Ragnar. 1934. Statistical confluence analysis by means of complete regression systems. Oslo.

28. Garrett, Henry E. 1943. The discriminant function and its use in psychology. *Psychometrika*, 8, 65-79.

29. Guilford, J. P. 1936 (a). The determination of item difficulty when chance success is a factor. *Psychometrika*, 1, 259-264.

30. ———. 1936. (b). Psychometric methods. New York: McGraw-Hill.

31. ———. 1937. The psychophysics of mental test difficulty. *Psychometrika*, 2, 121-133.

32. Gulliksen, Harold. 1936. The content reliability of a test. *Psychometrika*, 1, 189-194.

33. Hawkes, H. E., Lindquist, E. F., and Mann, C. R. 1936. The construction and use of achievement examinations. Boston: Houghton-Mifflin Company.

34. Hildreth, G. H. 1939. A bibliography of mental tests and rating scales. 2nd Ed. New York: The Psychological Corporation. Pp. xxiv + 295.

35. Holmes, Henry W. 1917. A descriptive bibliography of measurement in elementary subjects. Cambridge, Mass.: Harvard Univ. Press.

36. Holzinger, Karl J., and Clayton, Blythe. 1925. Further experiments in the application of Spearman's prophecy formula. *J. educ. Psychol.*, 16, 289-299.

37. Horst, Paul. 1934. (a). Item selection by the method of successive residuals. *J. exper. Educ.*, 2, 254-263.

38. ———. 1934 (b). Increasing the efficiency of selection tests. *The Personnel Journal*, 12, 254-259.

39. ———. 1936 (a). Obtaining a composite measure from different measures of the same attributes. *Psychometrika*, 1, 53-60.

40. ———. 1936 (b). Item selection by means of a maximizing function. *Psychometrika*, 1, 229-244.

41. Hull, Clark L. 1928. Aptitude testing. New York: World Book Company. Pp. xiv + 535.

42. Kelley, Truman L. 1927. Interpretation of educational measurements. New York: World Book Company.

43. ———. 1924. Statistical methods. New York: Macmillan Company. Pp. xi + 389.

44. ———. 1923. The principles and techniques of mental measurement. *Amer. J. Psychol.*, 34, 408-432.

45. Kuder, G. F. 1937. Nomograph for point biserial *r*, biserial *r*, and fourfold correlations. *Psychometrika*, 2, 135-138.

46. Kuder, G. F., and Richardson, M. W. 1937. The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160.

47. Lee, J. M., and Symonds, P. M. 1934. New type or objective tests: a sum-

mary of investigations (Oct. 1931-Oct. 1933). *J. educ. Psychol.*, 25, 161-184.

48. Lentz, T. F., Hirshstein, Bertha, and Finch, J. H. 1932. Evaluation of methods of evaluating test items. *J. educ. Psychol.*, 23, 344-350.
49. Lindquist, E. F. 1940. Statistical analysis in educational research. New York: Houghton Mifflin Co.
50. Long, John A., Sandiford, Peter, et al. 1935. The validation of test items. Bull. Dept. Educ. Res., Ontario Coll. Educ., No. 3, 126 pages.
51. McCall, W. A. 1922. How to measure in education. New York: The Macmillan Company. Pp. xii + 416.
52. Merrill, Walter W., Jr. 1937. Sampling theory in item analysis. *Psychometrika*, 2, 215-224.
53. Monroe, Paul (Editor). 1939. Conference on examinations at Dinard, France, Sept. 16-19, 1938. New York: Bureau of Publications, Teachers College, Columbia University. Pp. xiii + 330.
54. Monroe, Walter S. 1923. The theory of educational measurements. New York: Houghton-Mifflin Company.
55. Monroe, Walter S. 1934. A note on effiiciency of prediction. *J. educ. Psychol.*, 25, 547-548.
56. Moore, Clarence Carl. 1940. The rights-minus wrongs method of correcting chance factors in the T-F examination. *J. genet. Psychol.*, 57, 317-326.
57. Monroe, Walter S. and Englehart, Max D. 1936. Scientific study of educational problems. New York: The Macmillan Company.
58. Mosier, Charles I. 1936. A note on item analysis and the criterion of internal consistency. *Psychometrika*, 1, 275-282.
59. ———. 1940. Psychophysics and mental test theory: fundamental postulates and elementary theorems. *Psychol. Rev.*, 47, 355-366.
60. National Society for the Study of Education. 1918. 17th Yearbook, Part II. Bloomington, Ill.: Public School Publishing Company.
61. Orleans, Jacob S. 1937. Measurement in education. New York: Thomas Nelson and Sons. Pp. xvi + 461.
62. Otis, A. S. 1922 (a). The method for finding the correspondence between scores in two tests. *J. educ. Psychol.*, 13, 529-45.
63. ———. 1922 (b). A method of inferring the change in a coefficient of correlation resulting from a change in the heterogeneity of the group. *J. educ. Psychol.*, 13, 293-294.
64. Otis, A. S., and Knollin, H. E. 1921. The reliability of the Binet scale and of pedagogical scales. *J. educ. Research*, 4, 121-142.
65. Richardson, M. W. 1935. Abac for computing tetrachoric coefficients in item analysis. Chicago: Univ. Chicago Board of Examination.
66. ———. 1936 (a). Notes on the rationale of item analysis. *Psychometrika*, 1, 69-76.
67. ———. 1936 (b). The relation of difficulty to the differential validity of a test. *Psychometrika*, 1, 33-49.
68. Richardson, M. W., and Adkins, Dorothy C. 1938. A rapid method of selecting test items. *J. educ. Psychol.*, 29, 547-552.
69. Richardson, M. W., and Kuder, G. F. 1939. The computation of test reliability by the method of rational equivalence. *J. educ. Psychol.*, 30, 681-687.
70. Ruch, G. M. and Stoddard, G. P. 1927. Tests and measurements in high-school instruction. New York: World Book Company.
71. Ruch, G. M., Ackerson, L., and Jackson, J. P. 1926. An empirical study of the Spearman-Brown formula as applied to educational test material. *J. educ. Psychol.*, 17, 309-313.

72. Ruger, Georgie J. 1918. Bibliography of psychological tests. New York: Bureau of Educational Measurements.

73. Rulon, Phillip J. 1939. A simplified procedure for determining the reliability of a test by split halves. *Harvard educ. Rev.*, 9, 99-103.

74. Segel, David. 1933. A note of an error made in investigations of homogeneous grouping. *J. educ. Psychol.*, 24, 64-66.

75. Smith, B. O. 1938. Logical aspects of educational measurement. New York: Columbia Univ. Press.

76. Spearman, Charles. 1910. Correlation from faulty data. *Brit. J. Psychol.*, 3, 271-295.

77. ———. 1907. Demonstration of formulae for true measurement of correlation. *Amer. J. Psychol.*, 18, 161-169.

78. Stalnaker, J. M. 1938. Weighting questions in the essay-type examination. *J. educ. Psychol.*, 29, 481-490.

79. Stalnaker, J. M. and Richardson, M. W. 1933. A note concerning the combination of test scores. *J. gen. Psychol.*, 8, 460-463.

80. Starch, D., and Elliot, E. C. 1912. Reliability of grading high-school work in English. *School Review*, September, 442-457.

81. ———. 1913. Reliability of grading high-school work in mathematics. *School Review*, April, 254-259.

82. ———. 1913. Reliability of grading high-school work in history. *School Review*, December. 676-681.

83 Stern, William. 1914. The psychological methods of testing intelligence. Baltimore: Warwick and York.

84. Swineford, F. 1936. Validity of test items. *J. educ. Psychol.*, 27, 68-78.

85. Thorndike, E. L. 1922. On finding equivalent scores in tests of intelligence. *J. appl. Psychol.*, 6, 29-33.

86. Thurstone, L. L. 1919. A method for scoring tests. *Psychol. Bull.*, 16, 235-240.

87. ———. 1925. A method of scaling psychological and educational tests. *J. educ. Psychol*, 16, 433-451.

88. ———. 1926. The mental age concept. *Psychol. Rev.*, 33, 268-278.

89. ———. 1928. The absolute zero in intelligence measurement. *Psychol. Rev.*, 35. 175-197.

90. ———. 1931. The reliability and validity of tests. Ann Arbor, Mich.: Edwards Brothers. Planographed.

91. ———. 1924. Fundamentals of statistics. New York: The Macmillan Company. Pp. xvi + 237.

92. Thurstone, T. G. 1932. The difficulty of a test and its diagnostic value. *J. educ. Psychol.*, 23, 335-343.

93. Toops, H. A., and Symonds, P. M. 1922. What shall we expect of the A. Q.? *J. educ. Psychol.*, 13, 513-528.

94. Travers, R. M. W. 1939. The use of a discriminant function in the treatment of psychological group differences. *Psychometrika*, 4, 25-32.

95. Walker, Helen M. 1929. Studies in the history of statistical method. Baltimore: The Williams and Wilkins Company. Pp. 186.

96. Whipple, Guy M. 1910. Manual of mental and physical tests. Baltimore: Warwick and York. Vols. I and II.

97. Wilks, S. S. 1938. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3, 23-40.

98. Yerkes, R. M., Bridges, J. W., and Hardwick, R. S. 1915. A point scale for measuring mental ability. Baltimore: Warwick and York.

PSYCHOMETRIKA—VOL. 8, NO. 4
DECEMBER, 1943

THE CONCEPT OF TEST AND ITEM RELIABILITY IN RELATION TO FACTOR PATTERN

ROBERT J. WHERRY AND RICHARD H. GAYLORD
UNIVERSITY OF NORTH CAROLINA

It is shown that approaches other than the internal consistency method of estimating test reliability are either less satisfactory or lead to the same general results. The commonly attendant assumption of a single factor throughout the test items is challenged, however. The consideration of a test made up of K sub-tests each composed of a different orthogonal factor disclosed that the assumption of a single factor produced an erroneous estimate of reliability with a ratio of $(n-K)/(n-1)$ to the correct estimate. Special difficulties arising from this error in application of current techniques to short tests or to test batteries are discussed. Application of this same multi-factor concept to item-analysis discloses similar difficulties in that field. The item-test coefficient approaches $\sqrt{1/K}$ as an upper limit rather than 1.00 and approaches $\sqrt{1/n}$ as a lower limit rather than .00. This latter finding accounts for an over-estimation error in the Kuder-Richardson formula (8). A new method of isolating sub-tests based upon the item-test coefficient is proposed and tentatively outlined. Either this new method or a complete factor analysis is regarded as the only proper approach to the problem of test reliability, and the item-*sub-test* coefficient is similarly recommended as the proper approach for item analysis.

The term reliability has been used loosely to apply to the resultant of the application of many different statistical operations. As a general rule these operations attempt to determine the verifiability of the original data and thus to establish the degree to which non-chance factors entered into the original measurements.

One group of techniques approaches the problem through the direct comparison of the observed distribution of measurements with that which would have arisen by chance in accordance with certain theories of probability. These techniques parallel the use of the critical ratio in establishing the reliability of means, sigmas, differences, and other statistical parameters by disproving the null hypothesis. One such approach is that of Jackson (7), who measures the sensitivity of the test, gamma (γ), by taking the ratio of the standard deviation of the capacity to the standard deviation of chance. Another writer, Hoyt (6), using the analysis of variance approach, suggests using the ratio of the true variance of the student responses to the obtained variance among students. Edgerton and K. F. Thomson (3) suggest using the Lexis ratio to show that differences among students are greater than those among items. Hoyt shows that his results are comparable to those of Jackson, and both Hoyt and Edgerton and Thomson show their results to be comparable to those achieved by the use of certain of the Kuder-Richardson series [actually to formulas (14) and (20), which involve only the sigmas of the items and the sigma of the total test] which we will discuss later at some length. Any restrictions demonstrated to hold for these particular portions of the Kuder-Richardson series will thus apply to these probability methods as well. Allied techniques are (1) the Horst (4) maximized criterion which holds that the most reliable weighted composite is that with the largest relative variance, and (2) the item selection techniques based on the variance (difficulty) of the item. The remaining techniques attack the measurement of reliability through the verifiability of the original ranks of the members of the population. These techniques form two natural groups as the emphasis on stability is (1) regardless of time, or (2) regardless of the particular test or items used.

The first of these viewpoints—verifiability regardless of time—is exemplified by the test-retest method of measuring reliability. While appearing best to meet the operations indicated in the usual definition of reliability, this method has been widely criticised from many different viewpoints. Typical criticisms refer to the effects of differential practice, memory, inability to duplicate testing conditions, inability to sustain motivation, etc. Perhaps even more serious is the objection raised that this coefficient is affected not only by the unreliability of the test but also by the unreliability (lack of stability) of the function being tested. Paulsen (12) suggested measuring this trait fluctuation by correcting the test-retest coefficient for attenuation, using the split-half method for obtaining the reliabilities of the initial and final testings. Woodrow (18) suggested measuring this "quotidian variability" by the ratio of the actual sigma of the means of various samples to the sigma of the means as predicted from the average standard deviation of the multiple samples. Thouless (17) proposed the measurement of this "functional fluctuation" by what he called the double test-retest index, where he substituted alternate forms for the split-half approach of Paulsen. Thus in application we see that the "regardless of time criterion" in addition to many other ills requires an appeal to the preceding view—as in Woodrow—or to the remaining criterion of stability "regardless of test or items used" criterion—as in the split-half or comparable form methods of Paulsen or Thouless. Indirectly associated techniques are: (1) the original

Brown (1) concept as to the proper values for substitution in the Brown-Spearman formula; and (2) the concept of testing item reliability by means of individual changes in response on future test adminstrations.

The remaining concept—the verifiability of the measurements regardless of the particular test used—is exemplified by the "comparable test" method of measuring reliability. Here the procedure is complicated by two major difficulties: (1) the time element, and (2) the determination of *comparability*. With respect to time the argument centers in the merit of separate alternate tests (assuming "comparability" for the time being), with the necessary time separation involved, as compared to some split-half (odds vs. evens, first half-last half, etc.) technique (again assuming "comparability") which would eliminate the time element. All of the objections raised against the test-retest method with the possible exception of the memory element would also apply against any time-separated alternate form approach. Most recent writers have assumed the desirability of eliminating this time element by some method based on internal analysis.

The second difficulty involved in this last viewpoint—the nature and determination of "comparability"—has usually been dealt with inadequately. It is always assumed but rarely demonstrated. The only really adequate definition of such comparability or equivalence known to the writers is that given by Kuder and Richardson (9) as their equation (1) which we hereby adopt. They say,

"The correlation between two forms of a test is given by

$$r_{(a+b+c+\cdots+n)(A+B+C+\cdots+N)} = \frac{r_{aA}\,\sigma_a\,\sigma_A + r_{aB}\,\sigma_a\,\sigma_B + \cdots + r_{nM}\,\sigma_n\,\sigma_M + r_{nN}\,\sigma_n\,\sigma_N}{[\sigma_a^2 + \sigma_b^2 + \cdots + \sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + r_{ac}\,\sigma_a\,\sigma_c + \cdots + r_{mn}\,\sigma_m\,\sigma_n)]^{\frac{1}{2}}\;[\sigma_A^2 + \sigma_B^2 + \cdots + \sigma_N^2 + 2(r_{AB}\,\sigma_A\,\sigma_B + r_{AC}\,\sigma_A\,\sigma_C + \cdots + r_{MN}\,\sigma_M\,\sigma_N)]^{\frac{1}{2}}}, \quad (1)$$

in which a , b , $\cdots$, n are items of the test, and A , B , $\cdots$, N are corresponding items in a second hypothetical test. Equivalence is now defined as interchangeability of items a and A , b and B , etc.; the members of each pair have the same difficulty and are correlated to the extent of their respective reliabilities. The inter-item correlations of one test are the same as those in the other. These relationships constitute the operational definition of equivalence which is to be used."

As Kuder and Richardson point out, the above definition of comparability makes the two terms in the bottom of equation (1) identical, which reduces the formula for the true reliability of a test or test battery with unit weights to the form

$$r_{tt} = \frac{r_{aa}\,\sigma_a^2 + r_{bb}\,\sigma_b^2 + \cdots + r_{nn}\,\sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + \cdots + r_{mn}\,\sigma_m\,\sigma_n)}{\sigma_a^2 + \sigma_b^2 + \cdots + \sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + \cdots + r_{mn}\,\sigma_m\,\sigma_n)}. \qquad (2)$$

If such a test as the capital letter series above were available it would truly be an alternate or comparable form. Actually such a test is seldom if ever available. Attempts at constructing alternate forms seldom hold rigorously to the above definitions. Instead it has been the custom to make either explicitly or implicitly the assumption of a single factor running through all such possible items and to construct alternate forms paying attention only to the difficulty of the items (equating means and sigmas, if even these are taken into account).

The internal consistency hypothesis is the basis of the two most common methods of measuring reliability: (1) the split-half Spearman-Brown (15) approach, and (2) the Kuder-Richardson series. Kuder and Richardson frankly assume a single factor among the test items, while the Spearman-Brown assumption of equal intercorrelations amounts to the same thing as is demonstrated when Kuder and Richardson derive the Spearman-Brown formula at one stage in their own series. Allied techniques are (1) the Edgerton-Kolbe (2) conception of the most reliable weighted criterion based on minimal differences among the scores of each individual for each of the criteria; (2) the Hotelling (5) conception of the most predictable criterion with weights proportional to the loadings on the first unrotated principle component; (3) and the many methods of item analysis, too numerous even to list, based on the item-test correlation coefficient or variants of that measure. When the assumption of a single factor is satisfied such procedures are justified, but frequently that assumption is not justified as is indicated by the many studies which show the usual test or test battery to contain many factors. What is the result of using these formulas when, as is probably almost universally the case, the basic assumption is not justified? Have we, perhaps been branding as unreliable tests which were satisfactory in that respect? Have we been discarding as unreliable items which were perfectly good? The remainder of this paper considers these questions theoretically. Instead of the usual assumption of a single factor with equal intercorrelations and equal sigmas throughout the whole test, we shall instead assume a test of K factors—a , b , c , $\cdots$, k—each factor being represented by a number of items n_x which may vary from factor to factor. For the sake of simplicity and to maximize the difference between this case and the usual assumptions these factors shall be taken as orthogonal or uncorrelated, i.e., the intercorrelations between items in different factor groups are taken to be zero.* While the sigmas and inter-item correlations within a factor group may differ from group to group ($r_{aa} \neq r_{ff}$) and ($\sigma_a \neq \sigma_f$), we shall make the usual assumptions of equal sigmas and equal intercorrelations within each factor group ($r_{a_1a_5} = r_{a_3a_7}$ and $\sigma_{a_4} = \sigma_{a_8}$, etc.) and shall further assume the reliability of each item to be equal to the inter-item correlation holding for items in that group. Substituting these assumptions in equation (2) we see that the reliability of this multiple-factor test would be

$$r_{tt} = \frac{\sum^k n_x^2 \sigma_x^2 r_{xx}}{\sum^k n_x \sigma_x^2 + \sum^k n_x(n_x - 1) \ \sigma_x^2 r_{xx}}. \tag{3}$$

We want to compare with this value, the value as estimated by the single factor theory. The Kuder-Richardson formula (14) is taken as the best measure of this type since it is this formula which the various direct probability approaches equalled and which Kuder and Richardson have shown to be equal to the Brown-Spearman method which is basic to split-half approaches to the measurement of reliability. We shall also see later on that this is the most general of their formulas which is applicable, since formula (8) of their series is erroneous. The Kuder-Richardson formula (14) reads

$$r_{tt_{KR_{14}}} = \frac{\sigma_t^2 - \sum^n \sigma_x^2}{(\sum^n \sigma_x)^2 - \sum^n \sigma_x^2} \cdot \frac{(\sum^n \sigma_x)^2}{\sigma_t^2}. \tag{4}$$

Substituting our present assumptions for our *K*-factored test in this equation gives

$$r_{tt_{KR_{14}}} = \frac{\sum^K n_x(n_x - 1) \sigma_x^2 r_{xx}}{(\sum^K n_x \sigma_x)^2 - \sum^K n_x \sigma_x^2} \cdot \frac{(\sum^K n_x \sigma_x)^2}{\sum^K n_x \sigma_x^2 + \sum^K n_x(n_x - 1) \ \sigma_x^2 r_{xx}} \tag{5}$$

and to bring out the difference between equation (3) and (5) more clearly we can rewrite (5) as

* One of the editors objected to the stringency of this case, on the basis that it would seldom if ever occur in practice. We deliberately chose to magnify the discrepancy between our results and those based upon the assumption of a single factor, due to our feeling that the actual usual case would lie somewhere in between the two extremes. The special case formulas derived in this paper were not meant to be used computationally but only to disclose the effects of various possible trends which might exist in practical situations. We do believe that the possible sub-tests and their inter-relations form the only sound approach to either the problem of test reliability or the problem of item-analysis.

$$r_{tt_{KR14}} = \frac{(\sum^K n_x \sigma_x)^2}{(\sum^K n_x \sigma_x)^2 - \sum^K n_x \sigma_x^2} \left[\frac{\sum^K n_x^2 \sigma_x^2 r_{xx}}{\sum^K n_x \sigma_x^2 + \sum^K n_x (n_x - 1) \sigma_x^2 r_{xx}} - \frac{\sum^K n_x \sigma_x^2 r_{xx}}{\sum^K n_x \sigma_x^2 + \sum^K n_x (n_x - 1) \sigma_x^2 r_{xx}} \right] . \tag{5a}$$

Now the first term in this bracket is the true value as given by equation (3), which we see is reduced by an error factor (the second term in the bracket) and then increased again by a multiplier greater than one (the term preceding the bracket). We can see that the extent of this error will depend upon the actual values of the n's, σ's, and r's. To evaluate the extent of this error under certain special conditions we can simplify the formulas by assuming various ones of these determiners, n_x, σ, and r, to be equal from group to group. The following table gives the reduced form for r_{tt} and $r_{tt_{KR14}}$ for each such possible set of assumptions.

Several conclusions of interest and importance can be drawn on the basis of the equations in set (6). We note that:

(a) The Kuder-Richardson formula equals the true formula only when K, the number of factors, equals one (Case g, since for the remaining single set the n's, σ's and r's would all be equal, with K equal to one). With K equal to one, both formulas take the usual Brown-Spearman form, indicating that that formula is also correct for the usual assumption of a single factor.

(b) The Kuder-Richardson formula tends to underestimate the true reliability by the ratio $(n-K)/(n-1)$ when the number of factors, K, is greater than one. (Cases d, g, and k.)

(c) If every item in a test is perfectly reliable the test is perfectly reliable even though all intercorrelations are equal to zero (n equals K and all n_x's equal one) and regardless of the size of the sigmas (cases h, i, j, and k for the true r_{tt}) although the Kuder-Richardson would not indicate this fact, giving values all of the way down to zero for extreme cases.

(d) The size or uniformity of the sigmas is not important if the n's and r's are equal (Case e) since they then drop out of the formula. This indicates that they are the least important of the three determiners.

(e) The Brown-Spearman formula underestimates the true reliability by the ratio of $(n-K)/(n-1)$. (Case g.) The derivation of this formula is simple and easily understood. The usual Brown-Spearman formula reads

Equal	Unequal	r_{tt}^*	r_{ttKR14}	Case
n_x	σ, r	$\dfrac{n \sum r_{xx} \sigma_x^2}{K \sum \sigma_x^2 + (n-K) \sum r_{xx} \sigma_x^2}$	$\dfrac{n-K}{n - K \dfrac{\sum \sigma_x^2}{(\sum \sigma_x)^2}} \cdot r_{tt}$	a
σ	n_x, r	$\dfrac{\sum n_x^2 r_{xx}}{n + \sum n_x (n_x - 1) r_{xx}}$	$\dfrac{n}{n-1} \left[r_{tt} - \dfrac{\sum n_x r_{xx}}{n + \sum n_x (n_x - 1) r_{xx}} \right]$	b
r	σ, n_x	$\dfrac{r_{aa} \sum n_x^2 \sigma_x^2}{\sum n_x \sigma_x^2 + r_{aa} \sum n_x (n_x - 1) \sigma_x^2}$	$\dfrac{1}{1 - \dfrac{\sum n_x \sigma_x^2}{(\sum n_x \sigma_x)^2}} \left[r_{tt} - \dfrac{r_{aa} \sum n_x \sigma_x^2}{\sum n_x \sigma_x^2 + r_{aa} \sum n_x (n_x - 1) \sigma_x^2} \right]$	c
n_x, σ	r	$\dfrac{n \sum r_{xx}}{K^2 + (n-K) \sum r_{xx}}$	$\dfrac{n-K}{n-1} r_{tt}$	d
n_x, r	σ	$\dfrac{n\, r_{aa}}{K + (n-K) r_{aa}}$	$\dfrac{n-K}{n - K \dfrac{\sum \sigma_x^2}{(\sum \sigma_x)^2}} r_{tt}$	e
σ, r	n_x	$\dfrac{r_{aa} \sum n_x^2}{n + r_{aa} \sum n_x (n_x - 1)}$	$\dfrac{n}{n-1} \left[r_{tt} - \dfrac{r_{aa} \sum n_x}{n + r_{aa} \sum n_x (n_x - 1)} \right]$	f

(6)

Equal	Unequal	r_{tt}*	r_{ttKR14}	Case
n_x, r, σ	—	$\dfrac{n\, r_{aa}}{K + (n - K) r_{aa}}$	$\dfrac{n - K}{n - 1} r_{tt}$	g
$r = 1.00$	n_x, σ	$\dfrac{\sum n_x^2 \sigma_x^2}{\sum n_x \sigma_x^2 + \sum n_x (n_x - 1) \sigma_x^2} = 1.00$	$\dfrac{1}{1 - \dfrac{\sum n_x \sigma_x^2}{(\sum n_x \sigma_x)^2}} \left[1.00 - \dfrac{\sum n_x \sigma_x^2}{\sum n_x \sigma_x^2 + \sum n_x (n_x - 1) \sigma_x^2} \right]$	h
$r = 1.00, n_x$	σ	$\dfrac{n}{1 + (n - 1)} = 1.00$	$\dfrac{n - K}{n - K \dfrac{\sum \sigma_x^2}{(\sum \sigma_x)^2}} [1.00]$	i
$r = 1.00, \sigma$	n_x	$\dfrac{\sum n_x^2}{n + \sum n_x (n_x - 1)} = 1.00$	$\dfrac{n}{n - 1} \left[1.00 - \dfrac{n}{\sum n_x^2} \right]$	j
$r = 1.00, \sigma, n_x$	—	$\dfrac{n}{K + (n - K)} = 1.00$	$\dfrac{n - K}{n - 1} [1.00]$	k

* Note that when the n_x values are equal $n_x = n/K$, where n is the total number of items in the test.

$$r_{tt_{BS}} = \frac{n\, r_{xx}}{1 + (n-1)\, r_{xx}} \tag{7}$$

where r_{xx} equals the average intercorrelation for all the items, but for any given item under the condition of equal n's, σ's, and r's for each factor group the average intercorrelation would be

$$r_{xx} = \frac{(n_a - 1)\, r_{aa} + (n - n_a)\, 0}{n-1}, \tag{8}$$

and since n_a would equal n/K we would have

$$r_{xx} = \frac{n-K}{K}\, \frac{r_{aa}}{n-1}, \tag{9}$$

and substitution of this value in equation (7) yields

$$r_{tt_{BS}} = \frac{n-K}{n-1}\, \frac{n\, r_{aa}}{K + (n-K)\, r_{aa}}. \tag{6, g}$$

While this limiting error of $n-K/n-1$ for the two internal consistency methods of estimating reliability [see conclusions (b) and (e)] becomes negligible when n becomes very large, it is nevertheless true that in short tests or when the formulas are used to estimate the reliability of test batteries where the number of tests is usually small, the equations based on internal consistency (assumption of one factor) would lead to gross underestimation and serious theoretical difficulty.

As to the gross underestimation, we present two cases of 12-item tests or batteries with assumed intercorrelations within the factor groups (the reliability of each item) equal to .95 for one test and to .60 in the other. For each test we shall assume anywhere from one to twelve factors with from twelve to one items, respectively. The true and estimated values follow:

		$r_{aa} = .95$		$r_{aa} = .60$	
K	n_x	r_{tt}	r_{ttKR14}	r_{tt}	r_{ttKR14}
12	1	.950	.000	.600	.000
6	2	.974	.531	.750	.409
4	3	.983	.715	.818	.595
3	4	.987	.808	.857	.701
2	6	.991	.901	.900	.818
1	12	.996	.996	.947	.947

An example of attendant theoretical difficulty due to the erroneous use of the internal consistency theory of reliability is the supposed dilemma cited by Royer (14) for the problem of multiple correlation. Low intercorrelations are a prerequisite to achieving high validity, whereas the consistency hypothesis holds that low intercorrelations mean low reliability. This of course does not follow at all if we make the probably nearer correct assumption of multiple orthogonal factors in both the criterion and the test items. It is of course true, as Thompson (16) has shown, using the correct reliability formula, that maximum validity and maximum reliability are not identical goals. If the criterion happens to have its variance determined in large part by factors predominating among the less reliable portions of the possible predictor items, then maximum validity will mean greatly lowered reliability as compared with the prediction of a different criterion whose factor composition is proportional to the reliabilities of the selected items, i.e., with a comparable test. We cannot follow Thompson's idea of compromising the weights in this case, however, since getting a better measure of something a person does not want to measure—a more comparable test—is no advantage if one thereby secures a poorer measure of what he is trying to measure—the criterion. The proper mode of approach in the Thompson situation would be to improve the reliability, by lengthening or other means, of those portions of the test battery with high regression weights, thereby increasing validity as well as reliability, rather than the Thompson suggestion of shifting the weights so as to secure higher reliability at the expense of decreased validity.

One other serious consequence of the erroneous adoption of the single factor assumption is a widespread misuse of the item-test correlation coefficient as a method of item selection and elimination. It is true that this measure, as Richardson (13) has clearly demonstrated, is the appropriate measure to use when there is a single factor among the items, but let us examine the values of this coefficient for various items under the condition of multiple orthogonal factors assumed above. In general this coefficient for any given item will equal

$$r_{it} = \sigma_i[1 - (n_i - 1)r_{ii}]/\sqrt{\sum n_x \sigma_x^2 + \sum n_x(n_x - 1)\sigma_x^2 r_{xx}}\,, \tag{10}$$

where the subscript i stands for the particular group of items having the same factor as the item in question. We see then that the item-test coefficient depends not on reliability (r_{ii}) alone but also upon σ_i (item difficulty) and upon the number of items in the test measuring the same factor (n_i). Assuming these various determiners to be equal from group to group gives a series of values paralleling the various conditions in set (6) above:

Equal	Unequal	r_{it}	Case
n	σ, r	$\dfrac{\sigma_i[K + (n-K)r_{ii}]}{\sqrt{n}\sqrt{K\sum\sigma_x^2 + (n-K)\sum\sigma_x^2 r_{xx}}}$	a
σ	n, r	$\dfrac{1 + (n_i - 1)r_{ii}}{\sqrt{n + \sum n_x(n_x - 1)r_{xx}}}$	b
r	σ, n	$\dfrac{\sigma_i[1 + (n_i - 1)\bar{r}_{ii}]}{\sqrt{\sum n_x\sigma_x^2 + \bar{r}_{ii}\sum n_x(n_x - 1)\sigma_x^2}}$	c (11)
n, σ	r	$\dfrac{K + (n-K)r_{ii}}{\sqrt{n}\sqrt{K^2 + (n-K)\sum r_{xx}}}$	d
n, r	σ	$\sigma_i\sqrt{\dfrac{K + (n-K)r_{ii}}{n\sum\sigma_x^2}}$	e
σ, r	n	$\dfrac{1 + (n_i - 1)\bar{r}_{ii}}{\sqrt{(1 - \bar{r}_{ii})n + \sum n_x^2}}$	f
n, r, σ	—	$\sqrt{\dfrac{K + (n-K)\bar{r}_{ii}}{Kn}}$	g
$r = 1.00$	n, σ	$\dfrac{n_i\sigma_i}{\sqrt{\sum n_x^2\sigma_x^2}}$	h
$r = 1.00, n$	σ	$\sigma_i\sqrt{1/\sum\sigma_x^2}$	i
$r = 1.00, \sigma$	n	$n_i/\sqrt{\sum n_x^2}$	j
$r = 1.00, \sigma, n$	—	$\sqrt{1/K}$	k

From these equations in set (11) we can draw the following conclusions:

(a) The item-test coefficient is not a measure of item reliability alone, but depends upon the share of the total variance of the battery determined by the sub-battery of which it is a part.

(b) Perfectly reliable items with low n's and σ's would be discarded as worthless by the usual method of applying this criterion, while less reliable items with large n's and σ's would be retained. Items with the highest r_{it} value are not necessarily the best items.

(c) The value of r_{it} does not approach 1.00 as an upper limit as the items become more reliable, but instead, (n's and σ's being equal) approaches the value of $\sqrt{1/K}$. Of course if $K = 1$, r_{it} does approach 1.00, but if K is greater than one the possible upper limits are

	$r_{ii} = 1.00$	$r_{ii} = .50$
	Case k	Case g
K	upper limit of r_{tt}	
2	.707	.505
4	.500	.361
6	.408	.297
8	.354	.260
10	.316	.234
20	.224	.173

The automatic rejection of items with low r_{it} values is not justified.

(d) If r_{ii} is equal to zero, the lower limit of the value of r_{it} approaches $\sqrt{1/n}$ rather than zero, which makes the value of the Kuder-Richardson formula (8) (their article) fictitiously high. According to their concept of the single factor their formula (8) should, as their formula (14) does, become equal to zero when all of the intercorrelations are equal to zero, but this is not the case. Their formula (8) is

$$r_{tt_{KR8}} = \frac{\sigma_t^2 - \sum \sigma_i^2}{2\,\sigma_t^2} + \sqrt{\frac{\sum r_{it}^2 \sigma_i^2}{\sigma_t^2} + \left(\frac{\sigma_t^2 - \sum \sigma_i^2}{2\,\sigma_t^2}\right)^2}\,, \qquad (12)$$

which if we assume one factor with all σ's equal becomes

$$r_{tt_{KR8}} = \frac{\sigma_t^2 - n\,\sigma_i^2}{2\,\sigma_t^2} + \sqrt{\frac{\sigma_i^2 \sum r_{it}^2}{\sigma_t^2} + \left(\frac{\sigma_t^2 - n\,\sigma_i^2}{2\,\sigma_t^2}\right)^2}\,. \qquad (13)$$

If we let all r_{xx} values be 1.00, we have

$$\sigma_t^2 = n^2\,\sigma_i^2 \quad \text{and} \quad r_{it} = 1.00\,,$$

whence

$$r_{tt_{KR8}} = \frac{n^2\,\sigma_i^2 - n\,\sigma_i^2}{2\,n^2\,\sigma_i^2} + \sqrt{\frac{n\,\sigma_i^2}{n^2\,\sigma_i^2} + \left(\frac{n^2\,\sigma_i^2 - n\,\sigma_i^2}{2\,n^2\,\sigma_i^2}\right)^2} = 1.00\,, \tag{14}$$

which is the correct answer. But if we let all r_{xx} values equal .00 we have

$$\sigma_t^2 = n\,\sigma_i^2 \quad \text{and} \quad r_{it} = 1/\sqrt{n}$$

whence

$$r_{tt_{KR8}} = \frac{n\,\sigma_i^2 - n\,\sigma_i^2}{2\,n\,\sigma_i^2} + \sqrt{\frac{\sigma_i^2}{n\,\sigma_i^2} + \left[\frac{n\,\sigma_i^2 - n\,\sigma_i^2}{2\,n\,\sigma_i^2}\right]^2} = \sqrt{1/n} > .00, \tag{15}$$

which is incorrect. While this is the upper limit of error in this formula, it does remain spuriously large for all values of r_{xx} less than 1.00. The source of this spurious increment is the incorrect assumption of a self-correlation of 1.00 for r_{ii} in the item-test coefficient. While this is correct for the actual correlation with the test of which it is a part, the substitution of this value in the Kuder-Richardson series assumes it to hold as well for the "comparable" test series, the capital letter test, whereas the correct value here is not 1.00 but the actual reliability of the item. Thus when all r_{xx} values are unity the reliability is 1.00 and formula (8) is correct, but for all other values of r_{xx} the reliability will be less than unity and formula (8) will be in error, this error reaching a maximum as r_{xx} approaches zero.

Some numerical examples of typical solutions for 3-factor problems are presented here as an indication of the points made under conclusions (a), (b), and (c) above:

Case	n	σ	r	r_{it}	Case	n	σ	r	r_{it}
	50	.3	.90	.308		50	.5	.20	.105
a	50	.4	.70	.326	d	50	.5	.70	.303
	50	.5	.20	.141		50	.5	.90	.383
	100	.4	.20	.362		50	.5	.60	.559
b	40	.4	.70	.492	e	50	.4	.60	.447
	10	.4	.90	.158		50	.3	.60	.335
	100	.5	.60	.362		100	.5	.60	.534
c	40	.4	.60	.492	f	40	.5	.60	.216
	10	.3	.60	.158		10	.5	.60	.057

Case	n	σ	r	r_{it}	Case	n	σ	r	r_{it}
	100	.5	1.00	.940		100	.4	1.00	.887
h	40	.4	1.00	.304	j	40	.4	1.00	.355
	10	.3	1.00	.057		10	.4	1.00	.089
	50	.5	1.00	.421					
i	50	.4	1.00	.337					
	50	.3	1.00	.253					

In view of the general possible unsatisfactory condition resulting from the application of the present methods of estimating reliability of the total test and of item validation by the blind assumption of a single factor, it seems advisable to suggest a marked revision in the present methods of test analysis. The obvious solution would be the factorial analysis of each test, but while ideal this would be very laborious and often impracticable, especially if the number of items were at all large. The calculation of the intercorrelations would alone be a tremendous undertaking to say nothing of securing the residual matrices and the final rotation of the obtained loadings. Let us consider simpler possibilities.

A start toward such a method is found in the concept of "item synonymization" advanced by Lentz and Whitmer **(10)**. In their method item intercorrelations have to be computed and there are no clear-cut standards for inclusion of an item in any given synonymy. However, they have demonstrated that

(a) test items do tend to fall into groups,

(b) an item correlates more highly with its synonymy than with other synonymies, and

(c) synonymies tend to correlate lowly with each other.

If these synonymies were at all numerous in a test a much better estimate of its over-all reliability could be obtained by considering the synonymies as sub-tests, computing their reliabilities by the Kuder-Richardson formula(justified for such a consistent group), and then computing the total test reliability by using these coefficients together with the inter-synonymy correlations in their general formula (2). Item validity would be evaluated not in terms of the correlation with the total test put in terms of the item correlation with *any* of the synonymies.

The present writers suggest tentatively an approach based on item-test rather than inter-item correlation coefficients. The following steps constitute a job-analysis of the proposed method:

(1) Compute the total test score (T) using unit weights.

(2) Compute all item-test correlations (r_{iT}).

(3) Skim off the items with the highest such coefficients.

(4) Rescore the papers on the basis of the selected items (S_I).

(5) Compute all item correlations with the new score (r_{is_I}).

(6) Add new items where $r_{is_I} > r_{iT}$ and drop items where $r_{is_I} < r_{iT}$ to form S_I'.

(7) Rescore the papers on the basis of the amended list (S_I').

(8) Repeat steps (5) and (6), computing r_{is_I}', adding items where $r_{is_I}' > r_{is_I}$ and dropping them when $r_{is_I}' < r_{is_I}$.

(9) Compute S_I'' and repeat steps (7), (8), and (9) until no further changes are indicated, computing S_I''', S_I'''', $\cdots$, S_I^m.

(10) Record the final $r_{is_I}{}^m$ values and compute a new residual score $T' = T - S_I^m$.

(11) Repeat steps (2) through (9) using T' in place of T and S_{II}^x in place of S_I^x.

(12) Record the final $r_{is_{II}}{}^m$ values and compute a new residual test score $T'' = T' - S_{II}^m$ and continue as before.

(13) Repeat the entire process through T^m until the test items are exhausted or until all $r_{iT}{}^m$ values approximate zero.

The selected sub-tests S_I^m, S_{II}^m, S_{III}^m, etc., will correspond to the item synonymies of Lentz, will tend to lie along pre-rotated orthogonal axes (if the i items possessed simple structure to begin with), and the $r_{is_x}{}^m$ values will be the factor loadings of the individual items on those axes.

We can illustrate this method of taking the three-factor numerical examples given for case (f) in the discussion of r_{it}, above. Here the correlations of the items with the total test would be

100 values of .534,
40 values of .216,
and 10 values of .057.

We would select the 100 highest items as S_1, rescore the tests, and compute the r_{is_1} values, obtaining:

100 values of .602 ,*
and 50 values of .000 ,

completing that phase. We would next compute $T - S_1 = T'$ and compute $r_{iT'}$, obtaining

100 values of .000 ,
40 values of .589 ,
and 10 values of .155 .

We would then form S_2 from the 40 items with correlations of .589, rescore the papers, and then compute r_{is_2}, obtaining

110 values of .000 ,
and 40 values of .607 ,*

completing the second phase. We would next compute $T'' = T' - S_2$ and compute $r_{iT''}$, obtaining

140 values of .000 ,
10 values of .627 .*

Using these 10 values to form S_3, since rescoring would result in identical values, would complete the analysis of the test

The Lentz technique would have arrived at identical sub-tests but at a cost of 150 × 149/2 = 11,175 correlation coefficients rather than the present 6 × 150 = 900 sub-calculations. Of course all tests would not possess simple structure nor would they all contain an hierarchical arrangement such that $n_{s_1} > n_{s_2} > n_{s_3}$, etc. In case the several n's or the larger of these approached equality or in case the contribution to the variance approached equality the separation of the factors would become more laborious and perhaps impossible.

Mosier (11) has warned against the use of r_{it} in a two-factor test when (1) the factors are numerically equal, and (2) the items do not possess simple structure, i.e., have loadings on both factors. His general thesis is in line with the main argument of this paper, which extends the consideration to more than two factors and applies the idea to the whole problem of reliability as well as to item analysis. That his warning is pertinent as a criticism or possible limitation to the method of test analysis proposed above is also recognized.

* Computing the correlation of the item with the total not counting the item in question would produce the true value of .60 here and in later synonymies, but such error will not obscure the major relationships.

In two attempts to use this method empirically the writers found one test which worked out quite smoothly due to unequally represented factors, whereas in a second test the two main factors were equipotent and came out as a single sub-test until S_I was broken up into S_{IA} and S_{IB} by visual inspection, after which the method proceeded to work satisfactorily.

The present writers intend to examine further the application of the suggested method of analysis to both theoretical and empirical tests. If it can be made to overcome the difficulties involved in equipotent factors it may well provide a practicable means of approximate factorial analysis for test situations where the usual methods would be prohibitive. The results of such research will be published in a later paper.

An article by Kelley (8), published after the beginning of the present article, must be given special mention since it foreshadowed empirically a number of the rational conclusions presented above. His finding that formulas (14), (20), and (21) of the Kuder-Richardson series yielded, for a three-item test with inter-correlations of zero, values of .00 for the reliability coefficient, while their formula (8) yielded a value of .58, is a perfect example of our equation (15) showing the erroneous nature of their formula (8) and of the incorrectness of their general approach when the number of factors, K, is large in comparison with the number of items, n. Since the Kelley article did not point out the incorrectness of their formula (8) and especially since he repeated their claim that it was their most reliable equation, it was feared that his reporting of the value .58 might be taken as lending support to that spurious equation.

While we agree that the Kuder-Richardson series is a measure of coherence rather than of reliability, we cannot accept Kelley's coefficient of coherence, VC/SVC, since it gave a value of .33 for a test whose coherence is obviously zero. The reason for this erroneous result is the same as that responsible for the error in the Kuder-Richardson formula (8). They used unity for the item reliabilities while the Kelley method of factor analysis is equivalent to putting these same fictitious unities in the diagonals of the factor matrix. Had Kelley used the correct communalities of zero, i.e., had he used the Thurstone approach, he would have attained the correct coefficient of coherence of .00 as yielded by the Kuder-Richardson equations (14), (20), and (21). The fact that he obtained .33 while the Kuder-Richardson formula (8) yielded $\sqrt{.33}$ or .58 is empirical evidence of the similarity of error.

We also cannot agree to Kelley's appeal to an "act of judgment" on the part of the experimenter, as in splitting a test in half or other smaller fractions, as a valid or dependable method of computing reliability. The suggestion of a complete factor analysis as a basis for setting up sub-tests or the short sub-test selection method schematized above (if it proves practicable) with the then proper use of the basic Kuder-Richardson equation appears to be a sounder and much more scientific approach.

REFERENCES

1. Brown, W. Some experimental results in the correlation of mental abilities. *Brit. J. Psychol.*, 1909-1910, **3**, 296-322.
2. Edgerton, H. A., and Kolbe, L. E. The method of minimum variation for the combination of criteria. *Psychometrika*, 1936, **1**, 183-187.
3. Edgerton, H. A., and Thomson, K. F. Test scores examined with the lexis ratio. *Psychometrika*, 1942, **7**, 281-288.
4. Horst, P. Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, 1936, **1**, 53-60.
5. Hotelling, H. The most predictable criterion. *J. educ. Psychol.*, 1935, **26**, 139-142.
6. Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, **6**, 153-160.
7. Jackson, R. W. B. Reliability of mental tests. *Brit. J. Psychol.*, 1938-39, **29**, 267-287.
8. Kelley, T. L. The reliability coefficient. *Psychometrika*, 1942, **7**, 75-83.
9. Kuder, G. F., and Richardson, M. W. The theory of the estimation of reliability. *Psychometrika*, 1937, **2**, 151-160.
10. Lentz, T. F., and Whitmer, E. F. Item synonymization: a method for determining the total meaning of pencil-paper reactions. *Psychometrika*, 1941, **6**, 131-139.
11. Mosier, C. A note on item analysis and the criterion of internal consistency, *Psychometrika*, 1936, **1**, 275-282.
12. Paulsen, G. B. A coefficient of trait variability. *Psychol. Bull.*, 1931, **28**, 218.
13. Richardson, M. W. Note on the rationale of item analysis, *Psychometrika*, 1936, **1**, 69-76.
14. Royer, E. B. Some recent developments in test construction. *Proc. Okla. Acad. Sci.*, 1936, **16**, 107-109.
15. Spearman, C. Correlation calculated from faulty data. *Brit. J. Psychol.*, 1909-1910, **3**, 271-295.
16. Thompson, G. A. Weighting for battery reliability and prediction. *Brit. J. Psychol.*, 1939-40, **30**, 357-366.
17. Thouless, R. H. Test unreliability and functional fluctuation. *Brit. J. Psychol.*, 1935-36, **26**, 325-343.
18. Woodrow, H. Quotidian variability. *Psychol. Rev.*, 1932, **32**, 245-256.

PSYCHOMETRIKA—VOL. 8, NO. 4
DECEMBER, 1943

A NOMOGRAPH FOR RAPID DETERMINATION OF MEDIANS

CLIFFORD E. JURGENSEN
KIMBERLY-CLARK CORPORATION, NEENAH, WISCONSIN

Directions are given for constructing a very simple nomograph for computing medians, which is entered with information from the cumulative frequency distribution. It gives a linear interpolation within the class interval containing the median.

Computations of medians and semi-interquartile ranges are generally considered to be quickly and easily made. At times, and with regard to some types of work, such computations are so numerous, and consequently so time consuming, that the proposed project hardly seems worth the time required. This is perhaps especially true in the industrial situation where there is a critical shortage of clerical help and where expansion and turmoil resulting from the war situation have made the utilization of scientific techniques even more desirable than previously.

Such a situation confronted the author recently when developing an employee merit rating scale consisting of scaled statements which were to be checked by supervisors as applying or not applying to the employee being rated. A total of 352 statements indicative of employee merit were compiled, and each statement was given a rating which ranged from one to nine inclusive. One hundred supervisors rated each of the statements. In order to determine the scale value of each statement it was necessary to compute 352 medians, each being based on 100 cases. To determine whether supervisors agreed with each other regarding the value of each statement, and consequently whether or not it was usable in the final check list, it was necessary to compute 352 semi-interquartile ranges. For each of these, the first and third quartile had to be computed. Thus a total of 1,056 medians and quartiles had to be computed. Inasmuch as computations were being made by clerks relatively untrained in statistical procedures, it was decided to make each computation twice in order to assure accuracy, thus making a total of 2,112 required computations. Persons making the computations originally followed the usual formula of

$$\text{Median} = l + \left(\frac{\frac{N}{2} - F}{f_m} \right) i .$$

A similar formula, counting from the upper end of the distribution, was used as a check. The same formulas, changing the $N/2$, were used for the quartiles.

Shortly after the statistical computations had been started, it became obvious to the author that considerably more time would be required than had been anticipated. In the search for a shortcut, a graphical means for determining medians and quartiles was devised. The procedure is so simple that it is presented with apologies, although neither the author nor other psychologists with whom he has discussed it have seen or heard of the procedure being used previously. It has been found to give results identical with the arithmetical method, although more accurate because fewer errors were made than in the arithmetical method. No comparative record of time required was kept, but it is estimated that the graphical procedure required less than one-fourth the time necessary for the arithmetical method.

The nomograph was constructed merely by obtaining a large sheet of heavy quarter-inch cross-ruled paper and drawing two parallel lines 25 inches in length and ten inches apart. A cross bar was drawn halfway between the ends of each line, thus making a large H. The 25-inch length of the vertical lines allowed one quarter inch for each of the 100 cases in the frequency distribution. The horizontal line was divided into one-inch lengths, each inch representing one tenth of a point. (See Fig. 1 for reproduction in reduced size.)

The nomograph is used as follows: (1) the cumulative frequency of the distribution is obtained; (2) a rule is placed on the left vertical line at the point representing the cumulative frequency immediately below $N/2$; (3) the rule is pivoted on the left vertical line so that the right vertical line is crossed at the point representing the cumulative frequency immediately above $N/2$; (4) the number of tenths indicated by the point at which the rule crosses the horizontal line is added to the lower end of the interval in which the median falls. The resultant median is correct to the first decimal place, and may be estimated (if desired) to the second decimal place.

Original ratings of statements for which this nomograph was specifically devised, were made on a nine-point scale. Each point was considered to represent the range from that point plus or minus .5. In order to eliminate the necessity for adding tenths to a decimal number (e.g., .8 + 2.5), the one-inch sections on the cross bar were each increased in size by .5. The reading on the horizontal line was thus added directly to the midpoint score of the interval containing the cumulative frequency immediately below $N/2$. Inasmuch as medians were to be computed only to the first decimal place, sections on the horizontal line were marked to indicate the range included in each decimal number, e.g., the section marked .7 covers the range from .65 to .75.

First and third quartiles are computed in the same way as medians, the horizontal lines being drawn at $N/4$ and $3N/4$ respectively, these figures also being used in determining the cumulative frequencies at which each vertical line is entered. Obviously, any other percentile can be computed similarly.

Following is a specific example of the use of the nomograph (Fig. 1):

Ratings	1	2	3	4	5	6	7	8	9
Frequencies	1	8	23	46	17	4	1		
Cumulative Frequency	1	9	32	78	95	99	100		

To compute median:

1. Enter left vertical line at 32 (cumulative frequency just below $N/2$)
2. Enter right vertical line at 78 (c.f. just above $N/2$)
3. Read the tenths indicated where rule crosses horizontal line (.9) and add to midpoint of interval containing cumulative frequency just below $N/2$ (3.0) to secure median (3.9).

To compute first quartile:

1. Enter left vertical line at 9 (c.f. just below $N/4$)
2. Enter right line at 32 (c.f. just above $N/4$)
3. Read the tenths on the lower horizontal line (1.2) and add to midpoint of interval containing c.f. just below $N/4$ (2.0) to secure Q_1 (3.2).

To compute third quartile:

1. Enter left vertical line at 32 (c.f. just below $3N/4$)
2. Enter right vertical line at 78 (c.f. just below $3N/4$)
3. Read the tenths on upper horizontal line (1.4) and add to midpoint of interval containing c.f. just below $3N/4$ (3.0) to secure Q_3 (4.4).

As will be noted from the examples above, the points of entering the vertical lines for the quartiles are sometimes the same as for the median, with this example being the same for Q_3 as for the median. Considerable time can be saved if the user of the nomograph watches for such similarity.

The nomograph in Fig. 1 is published primarily for illustrative purposes and can be used only when N equals 100 or when N is expressed in terms of percentage. A similar nomograph can be constructed for any size N in less than ten minutes time. Cross-ruled paper should be used, and to insure accuracy each space on the vertical line should represent one individual. The distance between the vertical lines is immaterial provided the space can conveniently and accurately be divided into ten equal parts for results accurate to the first decimal place. Where desired, the distance between vertical lines can be divided into 100 spaces to secure accuracy correct to the second decimal place. Using eighth-inch cross-section paper, the distance would be only twelve and one half inches. Large nomographs of such size are easily read with great accuracy, and thus are superior to smaller reproductions such as illustrated here.

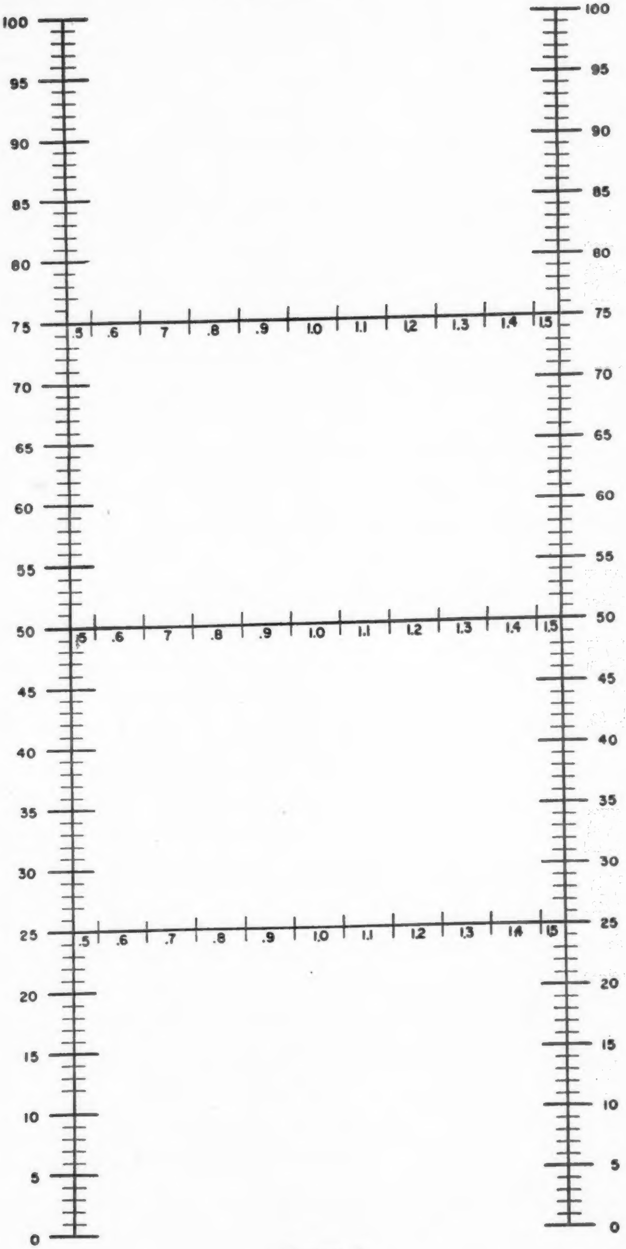
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
.5 .6 .7 .8 .9 1.0 1.1 1.2 1.3 1.4 1.5
.5 .6 .7 .8 .9 1.0 1.1 1.2 1.3 1.4 1.5
.5 .6 .7 .8 .9 1.0 1.1 1.2 1.3 1.4 1.5
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0

FIGURE 1

PSYCHOMETRIKA—VOL. 8, NO. 4
DECEMBER, 1943

INDEX FOR VOLUME 8

AUTHORS

Science

P

VOL
NUM
D

1 9

Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY · ORGANIZED IN 1935

VOLUME 8
NUMBER 4
DEC.
1943

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

DECEMBER 1943, VOLUME 8, NUMBER 4

Offices of Publication: 23 West Colorado Avenue, Colorado Springs, Colorado. Entered as second class matter, September 17, 1940 at the Post Office of Colorado Springs, Colorado, under the act of March 3, 1879. Editorial Office, 2539 Briarcliffe Avenue, Cincinnati, Ohio.

Subscription Price:

To non-members, the subscription price is $5.00 per volume of four issues. Members of the Psychometric Society pay annual dues of $5.00, of which $4.50 is in payment of a subscription to *Psychometrika.* Student members of the Psychometric Society pay annual dues of $3.00, of which $2.70 is in payment for the journal.

The subscription price to libraries and other institutions is $10.00 per year; this price includes one extra copy of each issue.

Applications for membership and student membership in the Psychometric Society should be sent to

THELMA GWINN THURSTONE
Chairman of the Membership Committee
University of Chicago
Chicago, Illinois

Membership dues and subscriptions are payable to

IRVING J. LORGE
Institute of Educational Research
Columbia University
New York, New York

(Continued on the back inside cover page)

Articles on the following subjects are published in *Psychometrika*:

(1) the development of quantitative rationale for the solution of psychological problems;

(2) general theoretical articles on quantitative methodology in the social and biological sciences;

(3) new mathematical and statistical techniques for the evaluation of psychological data;

(4) aids in the application of statistical techniques, such as monographs, tables, work-sheet layouts, forms, and apparatus;

(5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), in so far as articles of this type are available.

In the selection of the articles to be printed in *Psychometrika,* an effort is made to obtain objectivity of choice. All manuscripts are received by one person who first removes from each article the name of contributor and institution. The article is then sent to three or more persons who make independent judgments upon the suitability of the article submitted. This procedure seems to offer a possible mechanism for making judicious and fair selections.

Prospective authors are referred to the "Rules for Preparation of Manuscripts for *Psychometrika,*" contained in the September, 1939 issue. Reprints of these "Rules" are available upon request. A manuscript which fails to comply with these requirements will be returned to the author for revision.

Authors will be charged for cuts. They will receive 200 reprints without covers, free of charge.

Manuscripts for publication in *Psychometrika* should be sent to

HAROLD GULLIKSEN
Managing Editor Psychometrika
College Entrance Examination Board, Box 592
Princeton, New Jersey.

PRINTED BY THE DENTAN PRINTING CO., COLORADO SPRINGS, COLORADO